Evidence for information compression via the matching and unification of patterns in the workings of brains and nervous systems

J Gerard Wolff[*]

October 26, 2017

Abstract

This paper presents evidence for the idea that much of the workings of brains and nervous systems may be understood as information compression via the matching and unification of patterns (ICMUP). Information compression can mean selective advantage for any creature: in the efficient storage and transmission of information; and, owing to the close connection between information compression and concepts of prediction and probability, in the making of predictions about where food may be found, potential dangers, and so on. Several aspects of our everyday perceptions and thinking may be seen as information compression. For example, many words in natural languages may be seen as relatively short identifiers or 'codes' for relatively complex concepts. When viewing the world with two eyes, we see one view, not two. Random-dot stereograms provide confirmation that, in binocular vision, we do indeed merge information from our two eyes and thus compress it. Information compression may be seen in the workings of sensory units in the eye of *Limulus*, the horseshoe crab. Computer models demonstrate how information compression may be a key to the unsupervised discovery of grammars for natural language, including segmental structures (words and phrases), classes of structure, and abstract patterns. Information compression may be seen in the perceptual *constancies*, including size constancy, lightness constancy, and colour constancy. Mathematics, which is a product of the human intellect, may be seen to be a set of techniques for ICMUP, and their application. The *SP theory of intelligence*, with its empirical support, provides evidence for the importance of ICMUP, and more

[*] Dr Gerry Wolff BA (Cantab) PhD (Wales) CEng MIEEE MBCS; CognitionResearch.org, Menai Bridge, UK; jgw@cognitionresearch.org; +44 (0) 1248 712962; +44 (0) 7746 290775; *Skype*: gerry.wolff; *Web*: www.cognitionresearch.org.

specifically a concept of ‘SP-multiple-alignment’, in several aspects of human learning, perception, and thinking. Four objections to the main thesis of this paper are described, with answers to those objections.

Keywords: information compression, intelligence, perception, learning, cognition

1 Introduction

> “Fascinating idea! All that mental work I’ve done over the years, and what have I got to show for it? A goddamned zipfile! Well, why not, after all?” (John Winston Bush, 1996).

This paper describes observations and arguments that, in varying degrees, provide support for the idea that much of the workings of brains and nervous systems, including human learning, perception, and cognition, may be understood as information compression via the discovery of patterns that match each other and the merging or ‘unification’ of two or more instances of any pattern to make one. That perspective on information compression is described more fully in Section 2.4, below.

For the sake of brevity, the expression ‘human learning, perception, and cognition’ will be referred to as ‘HLPC’, the expression ‘information compression via the matching and unification of patterns’ will be referred to as ‘ICMUP’, and the main thesis of this paper—that much of the workings of brains and nervous systems may be understood as ICMUP—will be referred to as ‘BICMUP’.

The aim here is to review, update, and extend the discussion in [71], itself the basis for [72, Chapter 2], but with the main focus on the workings of brains and nervous systems.

The next section describes some of the background to this research and some relevant general principles, and the next-but-one section describes related research. Sections 4 to 15 inclusive describe empirical evidence in support of BICMUP, and Section 16, with Appendix B, describes apparent contradictions of ideas in this paper, and how they may be resolved.

2 Background and general principles

This section provides some background to this paper and summarises some general principles that have a bearing on BICMUP and the programme of research of which this paper is a part.

2.1 The importance of wide scope in theorising and in the creation and testing of computer models

As the title of this subsection suggests, this research attaches importance to maintaining a wide scope in the development of theories, and the importance of developing and testing computer models with wide scope as an aid to the development of theories.

In his famous essay, *You can't play 20 questions with nature and win*, Allen Newell [42] writes about the sterility of developing theories in narrow fields, and calls for each researcher to focus on "a genuine slab of human behaviour" (*ibid.*, p. 303).[1]

Newell's exhortation accords with a slightly extended version of Occam's Razor: in developing simple theories of empirical phenomena, we should concentrate on those with the greatest explanatory range. A theory that works well across a wide area is likely to be relatively robust and relatively immune to invalidation by new evidence.

In a similar vein, President Eisenhower is reputed to have said: "If you can't solve a problem, enlarge it", meaning that putting a problem in a broader context may make it easier to solve. Good solutions to a problem may be hard to see when the problem is viewed through a keyhole, but become visible when the door is opened.

With regard to the creation and testing of computer models with wide scope: this helps to guard against vagueness in theorising; it provides a very effective means of exposing the weaknesses of any idea; and it provides a means of demonstrating what a theory can do.

2.2 The *SP theory of intelligence* and its realisation in the *SP computer model*

In accordance with the principles outlined in Section 2.1, the *SP Theory of Intelligence* and its realisation in the *SP Computer Model* (outlined in Appendix C) is a unique attempt to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and HLPC, with information compression as a unifying theme.[2]

[1] Newell's essay and his book *Unified Theories of Cognition* [43] led to many attempts by himself and others to develop such theories. But in the light of Ben Goertzel's [24, p. 1] remark that "We have not discovered any one algorithm or approach capable of yielding the emergence of [general intelligence]." it seems that there is still some way to go.

[2] The name 'SP' is short for *Simplicity* and *Power*, because compression of any given body of information, **I**, may be seen as a process of reducing informational 'redundancy' in **I** and thus increasing its 'simplicity', whilst retaining as much as possible of its non-redundant

This paper relates to the SP research in two main ways:

- Since ICMUP is a central part of the SP theory, evidence for BICMUP presented in this paper in Sections 4 to 14 inclusive (but excluding Section 15) strengthens empirical support for the SP theory, viewed as a theory of HLPC.

- Empirical evidence for the SP theory as a theory of HLPC—summarised in Section 15—provides evidence for BICMUP which is additional to that in Sections 4 to 14 inclusive.

Owing to the wide scope of the SP theory and its many potential benefits and applications, it has proved difficult in writing about any one aspect of the system, or any one area of application, to avoid repeating information that has been presented elsewhere. In this paper and others, efforts have been made to avoid unnecessary repetition of information. At the same time, attention has been given to the need for clarity, and the need for each paper, including this one, to be free-standing so that it makes sense without recourse to other publications.

2.3 Approaches to information compression and probability

There are many approaches to information compression, many of them with a mathematical flavour (see, for example, [50]). Much the same is true of concepts of prediction and probability which, as outlined in Section 2.5, are closely related to information compression.

In the SP programme of research, the orientation is different. Amongst other things, the SP theory attempts to get below or behind the mathematics of other approaches to information compression and probability—to focus on ICMUP: the relatively simple, 'primitive' idea that information may be compressed by finding two or more patterns that match each other, and merging or 'unifying' them so that multiple instances of the pattern are reduced to one.[3] Of course, mathematics is very useful in many situations, including parts of the SP theory itself (see, for example, [72, Sections 3.5, 3.7, 3.10.6.2, and 9.2.6]). The intention in developing

expressive 'power'.

[3] An apparent exception to the generalisation that most approaches to information compression have a mathematical flavour is the widely-used 'LZ' algorithms [81, 82] and their variations (see, for example, "LZ77 and LZ78", *Wikipedia*, bit.ly/2wKOuhl, retrieved 2017-10-10). In these algorithms, the matching and unification of patterns is quite prominent. But the programs have little or no relevance to an understanding of HLPC because they are designed for speed on low-powered computers rather than the achievement of high levels of information compression, and of course they never aspired to be models of HLPC.

the SP system has been to avoid too much dependence on mathematics in the conceptual core of the theory.

There are three main reasons for this focus on ICMUP and the avoidance of too much dependence on mathematics:

- Since ICMUP is relatively 'concrete' and less abstract than the more mathematical approaches to information compression, it suggests avenues that may be explored in understanding possible mechanisms for compression of information and the estimation or calculation of probabilities, both in artificial systems and in brains and nervous systems. Here are two putative examples:
  - The concept of *SP-multiple-alignment* (Appendix C.1) is founded on ICMUP and is not a recognised part of today's mathematics—but it has proved to be effective in the compression of information, it makes possible a relatively straightforward approach to the calculation of probabilities for inferences, and it facilitates the modelling of several aspects of HLPC (Section 15, [80], [72, 74]).
  - The SP system, including the concept of SP-multiple-alignment with ICMUP, suggests how aspects of intelligence may be realised in a 'neural' version of the SP theory, *SP-neural*, expressed in terms of neurons and their interconnections (Appendix C.4).
- The SP theory, including ICMUP, aims to be, amongst other things, a theory of the foundations of mathematics [78], so it would not be appropriate for the theory to be too dependent on mathematics.
- Whilst the SP theory has benefitted from valuable insights gained from research on Algorithmic Probability Theory (APT), Algorithmic Information Theory (AIT), and related work (Section 3.2), it differs from that work in that it is *not* founded on the concept of a 'universal Turing machine' (UTM).

  Instead, a focus on ICMUP, has yielded *a new theory of computing and cognition*, founded on ICMUP and SP-multiple-alignment, with the generality of the UTM [72, Chapter 4] but with strengths in the modelling of human-like intelligence which are missing from the UTM ([80], [72, 74]).

Overall, the SP system, including ICMUP, provides a novel approach to concepts of information compression and probability which appears to have potential as an alternative to established methods in these areas.

A qualification to what has been said in this subsection is that, as argued in [78], mathematics may itself be seen to be founded on ICMUP. From that conclusion, it might be argued that, for the analysis of phenomena in HLPC, there is nothing to choose between ICMUP in mathematics and ICMUP in the SP system. In response to that argument: the SP system is a relatively direct expression of ICMUP; and it provides mechanisms such as SP-multiple alignment which are not provided in mathematics.

2.4 Six variants of ICMUP

In case the concept of ICMUP seems obscure, this subsection first describes 'basic ICMUP' more fully, and in Section 2.4.3, below, it describes five other variants which are widely used in everyday life and in science and engineering, and will be referred to later in the paper.

2.4.1 Searching for matches between patterns

The main idea in ICMUP is illustrated in the top part of Figure 1, below. Here, a stream of raw data may be seen to contain two instances of the pattern '`INFORMATION`'. Subjectively, we 'see' this immediately. But in a computer or a brain, the discovery of that kind of replication of patterns must necessarily be done by *a process of searching for matches between patterns.*

As indicated in Appendix C.1, the process of searching for matches between patterns can, in the process of building 'good' SP-multiple-alignments, be quite subtle and complex, including processes for finding good partial matches between patterns as well as full matches, as described in [72, Appendix A].

2.4.2 Unification of patterns

In itself, the detection of repeated patterns is not very useful. But by merging or 'unifying' the two instances of '`INFORMATION`' in Figure 1 we may create the single instance shown in the middle of the figure (excluding '`w62`' at the beginning). This kind of unification normally achieves a reduction in the overall size of the data. In other words, unification normally achieves compression of information.[4]

A discrete pattern like '`INFORMATION`' is often referred to as a *chunk* of information, a term that gained prominence in psychology largely because of its use by George Miller in his influential paper *The magical number seven, plus or minus two* [40].

Miller did not use terms like 'unification' or 'information compression', and he acknowledges some uncertainty about the significance of the concept of a chunk: "The contrast of the terms *bit* and *chunk* also serves to highlight the fact that we

[4] The qualifying word 'normally' is needed because, to achieve lossless compression of information, the repeated patterns that are to be unified must occur more frequently in the raw data than one would expect by chance (Section 2.4.5, [72, Section 2.2.8.3]).

are not very definite about what constitutes a chunk of information." (*ibid.*, p. 93, emphasis in the original). However, he describes how chunking of information may achieve something like compression of information: "... we must recognize the importance of grouping or organizing the input sequence into units or chunks. Since the memory span is a fixed number of chunks, we can increase the *number of bits of information that it contains* simply by building larger and larger chunks, each chunk containing more information than before." (*ibid.*, p. 93, emphasis in the original) and "... the dits and dahs are organized by learning into patterns and ... *as these larger chunks emerge* the amount of message that the operator can remember increases correspondingly." (*ibid.*, p. 93, emphasis in the original).

2.4.3 Five other variants of ICMUP

Apart from basic ICMUP, the previously-mentioned five other variants of ICMUP are these:

- *Chunking-with-codes.* With each unified *chunk* of information, give it a relatively short name, identifier, or *code*, and use that as a shorthand for the chunk of information wherever it occurs.

  The idea is illustrated in Figure 1, where, in the middle of the figure, the relatively short code or identifier '`w62`' is attached to a copy of the 'chunk' '`INFORMATION`' (that pairing of code and unified chunk would be stored separately from the body of data that is to be compressed). Then, under the heading "Compressed data" at the bottom of the figure, each of the two original instances of '`INFORMATION`' is replaced by the code '`w62`' yielding an overall compression of the original data.

  Examples of chunking-with-codes from this paper are the use of 'ICMUP' as a shorthand for "information compression via the matching and unification of patterns", and 'HLPC' as a shorthand for "human learning, perception, and cognition". Thus the use of 'ICMUP' to replace each occurrence of "information compression via the matching and unification of patterns" means that, in effect, those many instances have been unified and reduced to the single instance used in the definition of ICMUP.

- *Schema-plus-correction.* This variant is like chunking-with-codes but the unified chunk of information may have variations or 'corrections' on different occasions.

  An example from everyday life is a menu in a restaurant or cafe. This provides an overall framework, something like '`starter main_course pudding`' which may be seen as a chunk of information. Each of the three elements of the menu may be seen as a place where each customer may make a choice or 'correction' to the menu. For example, one customer may choose '`starter(soup) main_course(fish) pudding(apple_pie)`' while another customer may choose '`starter(salad) main_course(vegetable_hotpot) pudding(ice_cream)`', and so on.

- *Run-length coding.* This variant may be used with any sequence of two or more copies of a pattern where each copy except the first one follows immediately after the preceding copy. In that case, it is only necessary to record one copy of the pattern, with the number of copies, or with symbols or 'tags' to mark the start and end of the sequence.

  For example, a repeated pattern like:

  '`INFORMATIONINFORMATIONINFORMATIONINFORMATIONINFORMATION`'

  may be reduced to something like '`INFORMATION(×5)`' (where '×5' records the number of instances of '`INFORMATION`'). Alternatively, the sequence may be reduced to something like '`p INFORMATION ... #p`', meaning that "the pattern '`INFORMATION`' is repeated an unspecified number of times between the start and end symbols '`p ... #p`'".

  This is something like the instruction: "From the old oak tree keep walking until you see the river". Here, 'the old oak tree' marks the start of the repetition, 'keep walking' describes the repeated operation of putting one foot in front of the other, and 'until you see the river' marks the end of the repetition.

- *Class-inclusion hierarchy with inheritance of attributes.* Here, there is a hierarchy of classes and subclasses, with 'attributes' at each level. At every level except the top level, each subclass 'inherits' the attributes of all higher levels.

  In simplified form, the class '`vehicle`' contains sub-classes like '`road_vehicle`' and '`rail_vehicle`', the class '`road_vehicle`' contains sub-classes like '`bus`', '`lorry`', and '`car`', and so on. An attribute like '`contains_engine`' would be assigned to the top level ('`vehicle`') and would be inherited by all lower-level classes, thus avoiding the need to record that information repeatedly at all levels in the hierarchy. Likewise for attributes at lower levels.

  Of course there are many subtleties in the way people use class-inclusion hierarchies, such as cross-classification, 'polythetic' or 'family resemblance' concepts (in which no single attribute is necessarily present in every member of the given category and there need be no single attribute that is exclusive to that category [53]), and the ability to recognise that something belongs in a class despite errors of omission, commission, or substitution. The way in which the SP system can accommodate those kinds of subtleties is discussed in [72, Sections 2.3.2, 6.4.3, 12.2, and 13.4.6.2].

- *Part-whole hierarchy with inheritance of contexts.* This is like a class-inclusion hierarchy with inheritance of attributes except that the hierarchical structure represents the parts and subparts of some entity, and any given part inherits information about all the higher-level parts. In much the same way as with a class-inclusion hierarchy, a part-whole hierarchy promotes economy by sidestepping the need for each part of an entity to store full and explicit information about the higher-level structures of which it is a part.

  A simple example is the way that a '`person`' has parts like '`head`', '`body`', '`arms`', and '`legs`', while an arm may be divided into '`upper_arm`', '`forearm`', '`hand`', and so on. In a structure like this, inheritance means that if one hears that a given person has an injury to his or her hand, one can infer immediately that that person's arm has been injured, and indeed his or her whole '`person`'.

2.4.4 ICMUP and SP-multiple-alignment

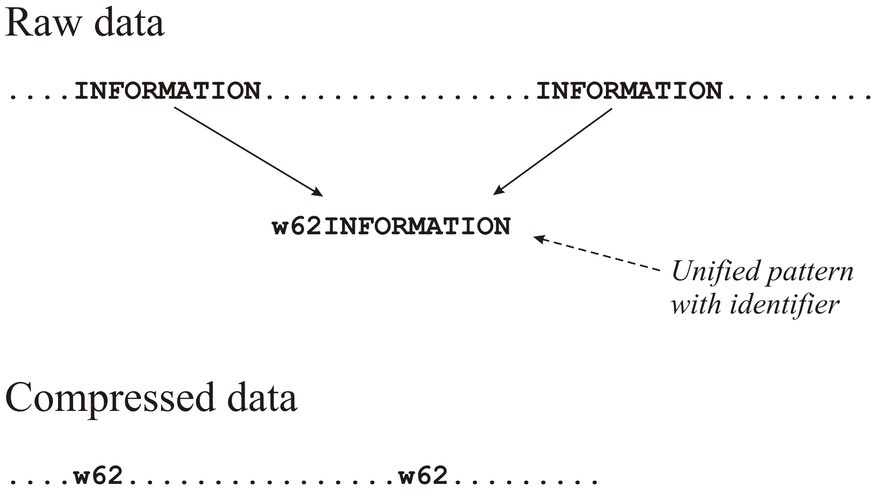
The way in which the six variants of ICMUP described in Section 2.4.3 may be realised via the concept of SP-multiple-alignment is described in [79, Appendix A.8]. In general, the SP-multiple-alignment construct provides for the seamless integration of these six variants of ICMUP, and perhaps others, in any combination.

Since the concept of SP-multiple-alignment also provides most of the versatility of the SP system, as outlined in Section 15, it may be seen as a super-charged version of ICMUP.

2.4.5 Quantification of information compression

So far, we have glossed over issues relating to the quantification of information compression in ICMUP. Here, in brief, are some key points:

- Any compression of information that may be achieved via basic ICMUP (Section 2.4) will be 'lossy', meaning that it will lose non-redundant information. This is because, in the unification of two or more matching patterns, information about the *location* of each pattern in the wider context is lost.



Raw data
....INFORMATION................INFORMATION.........
w62INFORMATION
Unified pattern with identifier
Compressed data
....w62................w62.........

Figure 1: A schematic representation of the way two instances of the pattern 'INFORMATION' in a body of raw data may be unified to form a single 'unified' pattern or 'chunk' of information, with 'w62' as a relatively short identifier or 'code' assigned by the system. The lower part of the figure shows how the raw data may be compressed by replacing each instance of 'INFORMATION' with a copy of the (shorter) identifer. Reproduced with permission from Figure 2.3 in [72].

- The chunking-with-codes technique is a means of avoiding the aforementioned loss of information about locations. This is because copies of the code associated with a unified pattern may be used to mark the locations of the patterns from which it was derived. This is illustrated in the lower part of Figure 1.

- With the chunking-with-codes technique, compression of information may be optimised by assigning shorter codes to more frequent chunks and longer codes to rarer chunks, in accordance with some such scheme as Shannon-Fano-Elias coding [17].

- With the other four techniques outlined in Section 2.4.3, similar principles may be applied.

- From the perspective of ICMUP, the concept of *redundancy* in information may be seen as the occurrence of two or more arrays of symbols that match each other, including arrays of symbols in which non-matching symbols are interspersed.

An important qualification here is that, for a given repeating array of symbols, **A**, to represent redundancy within a given body of information, **I**, **A**'s frequency of occurrence within **I** must be higher than would be expected by chance for an array of the same size [78, Appendix C].

2.4.6 Volumes of data and speeds of learning

An interesting corollary of the last point in the preceding subsection is that large patterns may exceed the threshold at a lower frequency than small patterns. With a complex pattern, such as an image of a person or a tree, there can be significant redundancy in a mere 2 occurrences of the pattern.

If redundancies can be detected via patterns that occur only 2 or 3 times in a given sample of data, unsupervised learning may prove to be effective with smallish amounts of data. This may help to explain why, in contrast to the very large amounts of data that are apparently required for success with deep learning, children and non-deep-learning types of learning program can do useful things with relatively tiny amounts of data [77, Section V-E].

In this connection, neuroscientist David Cox has been reported as saying: "To build a dog detector [with a deep learning system], you need to show the program thousands of things that are dogs and thousands that aren't dogs. My daughter only had to see one dog." and, the report says, she was happily pointing out puppies ever since.[5]

This issue relates to the way in which a camouflaged animal is likely to become visible when it moves relative to its background (Section 7). As with random-dot stereograms (Section 6), only two images which are similar but not the same are needed to reveal hidden structure.

2.5 Information compression and concepts of prediction and probability

It has been recognised for some time that there is an intimate relation between information compression and concepts of prediction and probability [63, 54, 55, 35].

In case this seems obscure, it makes sense in terms of ICMUP: a pattern that repeats is one that invites ICMUP, but it is also one that, via inductive reasoning, suggests what may happen in the future. As can be seen in the workings of the SP system, probabilities may be calculated from the frequencies with which different patterns occur ([72, Section 3,7], [74, Section 4.4]).

[5] "Inside the moonshot effort to finally figure out the brain", *MIT Technology Review*, 2017-10-12, bit.ly/2wRxsOg.

In more concrete terms, any repeating pattern—such as the association between black clouds and rain—provides a basis for prediction—black clouds suggest that rain may be on the way—and probabilities may be derived from the number of repetitions.

There is a little more detail in [78, Appendix D], and a lot more detail about how this works with the SP-multiple-alignment concept in [72, Section 3.7] and [74, Section 4.4].

The SP system has proved to be an effective alternative to Bayesian theory in explaining such phenomena as 'explaining away' ([72, Section 7.8], [74, Section 10.2]).

As indicated in Section 4, the close connection between information compression and concepts of prediction and probability makes sense in terms of biology.

2.6 Emotions and motivations

A point that deserves emphasis is that, while this paper is part of a programme of research aiming for simplification and integration of observations and ideas in cognitive psychology and related fields, it does not attempt to provide a comprehensive view of human psychology. In particular, it does not attempt to say anything about emotions or motivations, despite their undoubted importance and relevance to many aspects of human psychology, including cognitive psychology.

While the evidence is strong that ICMUP has an important role to play in HLPC, emotions and motivations are certainly important too.

3 Related research

An early example of thinking relating to information compression in cognition was the suggestion by William of Ockham in the 14th century that "Entities are not to be multiplied beyond necessity.". Later, there have been remarks by prominent scientists about the importance of simplicity in science (summarised in [78, Section 3]). Then research with a more direct bearing on BICMUP began in the 1950s and '60s after the publication of Claude Shannon's [52] 'theory of communication' (later called 'information theory'), and partly inspired by it.

In the two subsections that follow, there is a rough distinction between research with the main focus on issues in cognitive psychology and neuroscience, and research that concentrates on issues in mathematics and computing. In both sections, research is described roughly in the order in which it was published.

In this research, the prevailing view of information, and compression of information, is that they are things to be defined and analysed in mathematical terms. This perspective has yielded some useful insights, such as, for example, Solomonoff's demonstration of the close relation between information compression and concepts of prediction and probability (Section 2.5). But, as suggested in Section 2.3, there are potential advantages in the ICMUP perspective adopted in the SP research.

3.1 Psychology- and neuroscience-related research

Research relating to information compression and HLPC may be divided roughly into two parts: early research initiated in the 1950s and 60s by Fred Attneave, Horace Barlow and others, and then after a relative lull in activity, later research from the 1990s onwards.

3.1.1 Early research

In a paper called "Some informational aspects of visual perception", Fred Attneave [2] describes evidence that visual perception may be understood in terms of the distinction between areas in a visual image where there is much redundancy, and boundaries between those areas where non-redundant information is concentrated:

> "... information is concentrated along contours (i.e., regions where color changes abruptly), and is further concentrated at those points on a contour at which its direction changes most rapidly (i.e., at angles or peaks of curvature)." [2, p. 184].

For those reasons, he suggests that:

> "Common objects may be represented with great economy, and fairly striking fidelity, by copying the points at which their contours change direction maximally, and then connecting these points appropriately with a straight edge." [2, p. 185].

And he illustrates the point with a drawing of a sleeping cat reproduced in Figure 2.

Satosi Watanabe picked up the baton in a paper, referenced in Section 2.5, called "Information-theoretical aspects of inductive and deductive inference" [63]. He later wrote about the role of information compression in pattern recognition [64, 65].

At about this time, Horace Barlow published a paper called "Sensory mechanisms, the reduction of redundancy, and intelligence" [3] in which he argued, on the strength of the large amounts of sensory information being fed into the [mammalian] central nervous system, that "the storage and utilization of this enormous

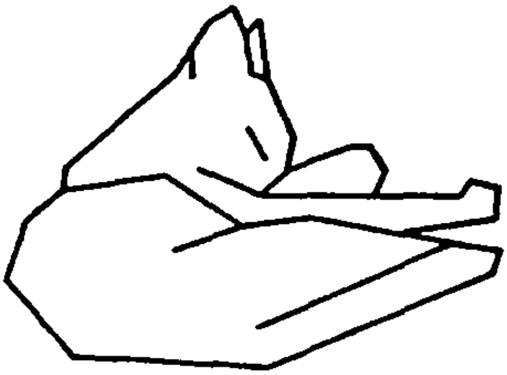


Figure 2: Drawing made by abstracting 38 points of maximum curvature from the contours of a sleeping cat, and connecting these points appropriately with a straight edge. Reproduced from Figure 3 in [2], with permission.

sensory inflow would be made easier if the redundancy of the incoming messages was reduced.” (*ibid.* p. 537).

In the paper, Barlow makes the interesting suggestion that:

> “... the mechanism that organises [the large size of the sensory inflow] must play an important part in the production of intelligent behaviour.” (*ibid.* p. 555).

and in a later paper he writes:

> “... the operations required to find a less redundant code have a rather fascinating similarity to the task of answering an intelligence test, finding an appropriate scientific concept, or other exercises in the use of inductive reasoning. Thus, redundancy reduction may lead one towards understanding something about the organization of memory and intelligence, as well as pattern recognition and discrimination.” [4, p. 210].

These prescient insights into the significance of information compression for the workings of human intelligence, with further discussion in [5], is a strand of thinking that has carried through into the SP theory of intelligence, with a wealth of supporting evidence, summarised in Section 15.[6]

Barlow developed these and related ideas over a period of years in several papers, some of which are referenced in this paper. However, in [6], he adopted a new position, arguing that:

[6] When I was an undergraduate at Cambridge University, it was fascinating lectures by Horace Barlow about the significance of information compression in the workings of brains and nervous systems, that first got me interested in those ideas.

> "... the [compression] idea was right in drawing attention to the importance of redundancy in sensory messages because this can often lead to crucially important knowledge of the environment, but it was wrong in emphasizing the main technical use for redundancy, which is compressive coding. The idea points to the enormous importance of estimating probabilities for almost everything the brain does, from determining what is redundant to fuelling Bayesian calculations of near optimal courses of action in a complicated world." (*ibid.* p. 242).

While there are some valid points in what Barlow says in support of his new position, his overall conclusions appear to be wrong. His main arguments are summarised in Appendix A, with what I'm sorry to say are my critical comments after each one.[7]

3.1.2 Later research

Later studies relating to information compression in brains and nervous systems have little to say about ICMUP. But they help to confirm the importance of information compression in HPLC, and thus provide some indirect support for BICMUP. A selection of publications are described briefly here.

- Ruma Falk and Clifford Konold [19] describe the results of experiments indicating that the perceived randomness of a sequence is better predicted by various measures of its encoding difficulty than by its objective randomness. They suggest that judging the extent of a sequence's randomness is based on an attempt to encode it mentally, and that the subjective experience of randomness may result when that kind of attempt fails.

- Jose Hernández-Orallo and Neus Minaya-Collado [27] propose a definition of intelligence in terms of information compression. At the most abstract level, it chimes with remarks by Horace Barlow quoted in Section 3.1.1, and indeed it is consonant with the SP theory itself. But the proposal shows no hint of how to model the kinds of capabilities that one would expect to see in any artificial system that aspires to human-like intelligence.

- Nick Chater, with others, has conducted extensive research on HLPC, compression of information, and concepts of probability, generally with an orientation towards AIT, Bayesian theory, and related ideas. For example:

[7] I feel apologetic about this because, as I mentioned, Barlow's lectures and his earlier research relating to information compression in brains and nervous systems have been an inspiration for me over many years.

  - Chater [9] discusses how 'simplicity' and 'likelihood' principles for perceptual organisation may be reconciled, with the conclusion that they are equivalent. He suggests that "the fundamental question is whether, or to what extent, perceptual organization is maximizing simplicity and maximizing likelihood." (*ibid.*, p. 579).
  - Chater [10] discusses the idea that the cognitive system imposes patterns on the world according to a simplicity principle, meaning that it chooses the pattern that provides the briefest representation of the available information. Here, the word 'pattern' means essentially a theory or system of one or more rules, a meaning which is quite different from the meaning of 'pattern' or 'SP-pattern' in the SP research, which simply means an array of atomic symbols in one or two dimensions. There is further discussion in [12].
  - Emmanuel Pothos with Nick Chater [45] present experimental evidence in support of the idea that, in sorting novel items into categories, people prefer the categories that provide the simplest encoding of these items.
  - Nick Chater with Paul Vitányi [13] describe how the 'simplicity principle' (information compression) allows the learning of language from positive evidence alone, given quite weak assumptions, in contrast to results on language learnability in the limit [25]. There is further discussion in [28]. These issues relate to discussion in Section 12.
  - Editors Nick Chater and Mike Oaksford [11] present a variety of studies using Bayesian analysis to understand probabilistic phenomena in HLPC.
  - Paul Vitányi with Nick Chater [59] discuss whether it is possible to infer a probabilistic model of the world from a sample of data from the world and, via arguments relating to AIT, they reach positive conclusions.

- Jacob Feldman [20] describes experimental evidence that, when people are asked to learn 'Boolean concepts', meaning categories defined by logical rules, the subjective difficulty of learning a concept is directly proportional to its 'compressibility', meaning the length of the shortest logically equivalent formula.

- Don Donderi [18] presents a review of concepts that relate to the concept of 'visual complexity'. These include Gestalt psychology, Neural Circuit Theory, AIT, and Perceptual Learning Theory. The paper includes discussion of how these and related ideas may contribute to an understanding of human performance with visual displays.

- Vivien Robinet and co-workers [49] describe a dynamic hierarchical chunking mechanism, similar to the MK10 computer model (Section 10). The theoretical orientation of this research is towards AIT, while the MK10 computer model embodies ICMUP.

- From analysis and experimentation, Nicolas Gauvrit and others [23] conclude that how people perceive complexity in images seems to be partly shaped by the statistics of natural scenes. [22] describe how it is possible to overcome the apparent shortcoming of AIT in estimating the complexity of short strings of symbols, and they show how the method may be applied to examples from psychology.

- In a review of research on the evolution of natural language, Simon Kirby and others [32] describe evidence that transmission of language from one person to another has the effect of developing structure in language, where 'structure' may be equated with compressibility. On the strength of further research, [57] conclude that increases in compressibility arise from learning processes (storing patterns in memory), whereas reproducing patterns leads to random variations in language.

- On the strength of a theoretical framework, an experiment and a simulation, Benoît Lemaire and co-workers [34] argue that the capacity of the human working memory may be better expressed as a quantity of information rather than a fixed number of chunks.

- In related work, Fabien Mathy and Jacob Feldman [39] redefine George Miller's [40] concept of a 'chunk' in terms of AIT as a unit in a "maximally compressed code". On the strength of experimental evidence, they suggest that the true limit on short-term memory is about 3 or 4 distinct chunks, equivalent to about 7 uncompressed items (of average compressibility), consistent with George Miller's famous magical number.

- And Mustapha Chekaf and co-workers [14] describe evidence that people can store more information in their immediate memory if it is 'compressible' (meaning that it conforms to a rule such as "all numbers between 2 and 6") than if it is not compressible. They draw the more general conclusion that immediate memory is the starting place for compressive recoding of information.

3.2 Mathematics- and computer-related research

Other research, with an emphasis on issues in mathematics and computing, can be helpful in the understanding of information compression in brains and nervous systems. This includes:

- Ray Solomonoff developed a formal theory known as *Algorithmic Probability Theory* showing the intimate relation between information compression and inductive inference [54, 55] (Section 2.5).

- Chris Wallace with others explored the significance of information compression in classification and related areas (see, for example, [61, 62, 1].

- Gregory Chaitin and Andrei Kolmogorov, working independently, built on the work of Ray Solomonoff in developing AIT. The main idea here is that the information content of a string of symbols is equivalent to the length of the shortest computer program that anyone has been able to devise that describes the string.

- Jorma Rissanen has developed related ideas in [47, 48] and other publications.

A detailed description of these and related bodies of research may be found in [35].

In research on deep learning in artificial neural networks, well reviewed by Jürgen Schmidhuber [51], there is some recognition of the importance of information compression (*ibid.*, Sections 4.2, 4.4, and 5.6.3), but it appears that the idea is not well developed in that area.

Marcus Hutter, with others, [29, 30, 58] has developed the 'AIXI' model of intelligence based on Algorithmic Probability Theory and Sequential Decision Theory. He has also initiated the 'Hutter Prize', a competition with €50,000 of prize money, for lossless compression of a given sample of text. The competition is motivated by the idea that "being able to compress well is closely related to acting intelligently, thus reducing the slippery concept of intelligence to hard file size numbers."[8] This is an interesting project which may yet lead to general, human-level AI.

4 Information compression and biology

This section and those that follow describe evidence that, in varying degrees, lends support to the BICMUP perspective.

First, let's take an abstract, bird's eye view of why information compression might be important in people and other animals. In terms of biology, information compression can confer a selective advantage to any creature:

[8] From www.hutter1.net, retrieved 2017-10-10.

- By allowing it to store more information in a given storage space or use less storage space for a given amount of information, and by speeding up the transmission of any given volume of information along nerve fibres—thus speeding up reactions—or reducing the bandwidth needed for the transmission of the same volume of information in a given time.

  In connection with the last point, we have seen in Section 3.1 how Barlow [3, p. 548] draws attention to evidence that, in mammals at least, each optic nerve is far too small to carry reasonable amounts of the information impinging on the retina unless there is considerable compression of that information.

- Perhaps more important than the impact of information compression on the storage or transmission of information is the close connection, outlined in Section 2.5 and noted in Section 3.2, between information compression and concepts of prediction and probability. Compression of information provides a means of predicting the future from the past and estimating probabilities so that, for example, an animal may learn to anticipate where food may be found or where there may be dangers.

  As mentioned in Section 2.5, the close connection between information compression and concepts of prediction and probability makes sense in terms of ICMUP: any repeating pattern can be a basis for predictions, and the probabilities of such predictions may be derived from the number of repetitions of the given pattern.

  Being able to make predictions and estimate probabilities can mean large savings in the use of energy with consequent benefits in terms of survival.

5 Hiding in plain sight

ICMUP is so much embedded in our thinking, and seems so natural and obvious, that it is easily overlooked. The following subsections describe some examples.

5.1 Chunking-with-codes

In the same way that ‘TFEU’ may be a convenient code or shorthand for the rather cumbersome expression ‘Treaty on the Functioning of the European Union’ (Appendix B.1.1), a name like ‘New York’ is a compact way of referring to the many things of which that renowned city is composed. Likewise for the many other names that we use: ‘Nelson Mandela’, ‘George Washington’, ‘Mount Everest’, and so on.

The ‘chunking-with-codes’ variant of ICMUP (Section 2.4.3) permeates our use of natural language, both in its surface forms and in the way in which surface forms relate to meanings.[9]

Because of its prominence in natural language and because of its intrinsic power, chunking-with-codes is probably important in other aspects of our thinking, as may be inferred from the way people naturally adopt this way of thinking, and, indirectly, via empirical support for the SP system (Section 15).

5.2 Class-inclusion hierarchies

In a similar way, class-inclusion hierarchies, with variations such as cross-classification, are prominent in our use of language and in our thinking, with consequent benefits arising from economies in the storage of information and in inferences via inheritance of attributes, in accordance with the ‘class-inclusion hierarchies’ variant of ICMUP (Section 2.4.3).

As with chunking-with-codes, names for classes of things provide for great economies in our use of language: most ‘content’ words in our everyday language stand for *classes* of things and, as such, are powerful aids to economical description. Imagine how cumbersome things would be if, on each occasion that we wanted to refer to a “table”, we had to say something like “A horizontal platform, often made of wood, used as a support for things like food, normally with four legs but sometimes three, ...”, like the slow *Entish* language of the Ents in Tolkien’s *The Lord of the Rings*.[10] Likewise for verbs like “speak” or “dance”, adjectives like “artistic” or “exuberant”, and adverbs like “quickly” or “carefully”.

5.3 Schema-plus-correction, run-length coding, and part-whole hierarchies

Again, it seems natural and obvious to conceptualise things in terms of other techniques mentioned in Section 2.4.3: schema-plus-correction, run-length coding, and part-whole hierarchies. And, as with chunking-with-codes and class-inclusion hierarchies, there are likely to be substantial benefits in terms of compression of information and the making of inferences.

[9] Although natural language provides a very effective means of compressing information about the world, it is not free of redundancy. And that redundancy has a useful role to play in, for example, enabling us to understand speech in noisy conditions, and in learning the structure of language (Appendix B.2 and [74, Section 5.2]).

[10] J. R. R. Tolkien, *The Lord of the Rings*, London: HarperCollins, 2005, Kindle edition. For a description of Entish, see, for example, page 480. See also, pages 465, 468, 473, 477, 478, 486, and 565.

5.4 Merging multiple views to make one

Here is another example of something that is so familiar that we are normally not aware that it is part of our perceptions and thinking.

If, when we are looking at something, we close our eyes for a moment and open them again, what do we see? Normally, it is the same as what we saw before. But recognising that the before and after views are the same, means unifying the two patterns to make one and thus compressing the information, as shown schematically in Figure 3.

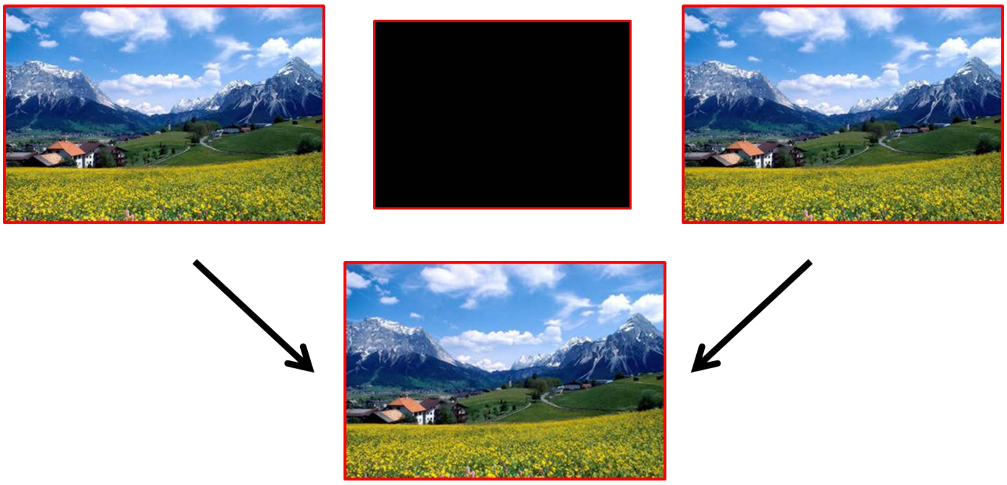


Figure 3: A schematic view of how, if we close our eyes for a moment and open them again, we normally merge the before and after views to make one. The landscape here and in Figure 4 is from Wallpapers Buzz (www.wallpapersbuzz.com), reproduced with permission.

It seems so simple and obvious that if we are looking at a landscape like the one in the figure, there is just one landscape even though we may look at it two, three, or more times. But if we did not unify successive views we would be like an old-style cine camera that simply records a sequence of frames, without any kind of analysis or understanding that, very often, successive frames are identical or nearly so.

5.5 Recognition

With the kind of merging of views just described, we do not bother to give it a name. But if the interval between one view and the next is hours, months, or years, it seems appropriate to call it 'recognition'. In cases like that, it is more obvious that we are relying on memory, as shown schematically in Figure 4.

Notwithstanding the undoubted complexities and subtleties in how we recognise things, the process may be seen in broad terms as ICMUP: matching incoming information with stored knowledge, merging or unifying patterns that are the same, and thus compressing the information.

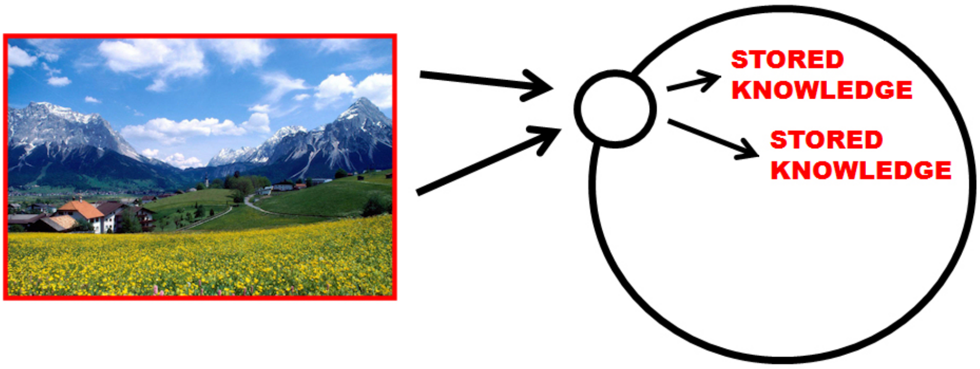


Figure 4: Schematic representation of how, in recognition, incoming visual information may be matched and unified with stored knowledge.

If we did not compress information in that way, our brains would quickly become cluttered with millions of copies of things that we see around us—people, furniture, cups, trees, and so on—and likewise for sounds and other sensory inputs.

As mentioned earlier, Satosi Watanabe has explored the relationship between pattern recognition and information compression [64, 65].

6 Binocular vision

ICMUP may also be seen at work in binocular vision:

> "In an animal in which the visual fields of the two eyes overlap extensively, as in the cat, monkey, and man, one obvious type of redundancy in the messages reaching the brain is the very nearly exact reduplication of one eye's message by the other eye." [4, p. 213].

In viewing a scene with two eyes, we normally see one view and not two. This suggests that there is a matching and unification of patterns, with a corresponding compression of information. A sceptic might say, somewhat implausibly, that the one view that we see comes from only one eye. But that sceptical view is undermined by the fact that, normally, the one view shows depth with a vividness that comes from merging the two slightly different views from both eyes.

Strong evidence that, in stereoscopic vision, we do indeed merge the views from both eyes, comes from a demonstration with 'random-dot stereograms', as described in [75, Section 5.1].

In brief, each of the two images shown in Figure 5 is a random array of black and white pixels, with no discernable structure, but they are related to each other as shown in Figure 6: both images are the same except that a square area near the middle of the left image is further to the left in the right image.

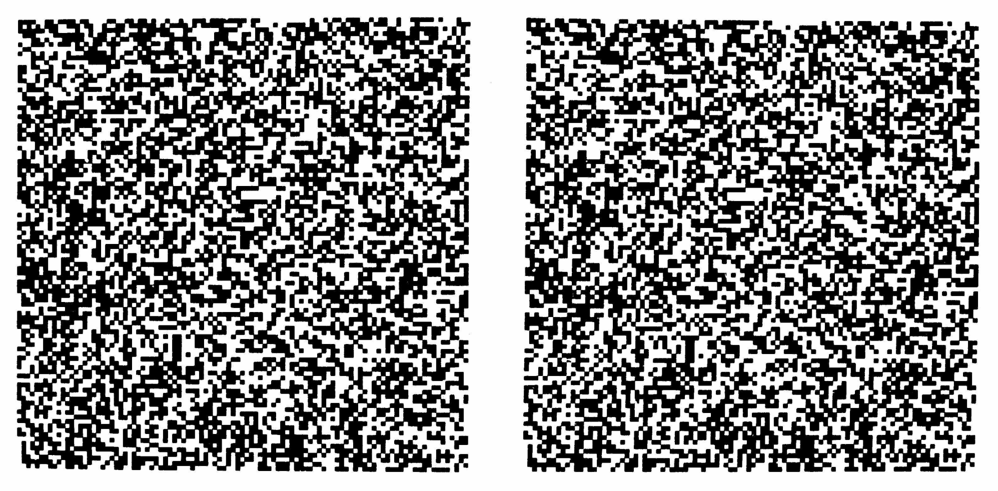


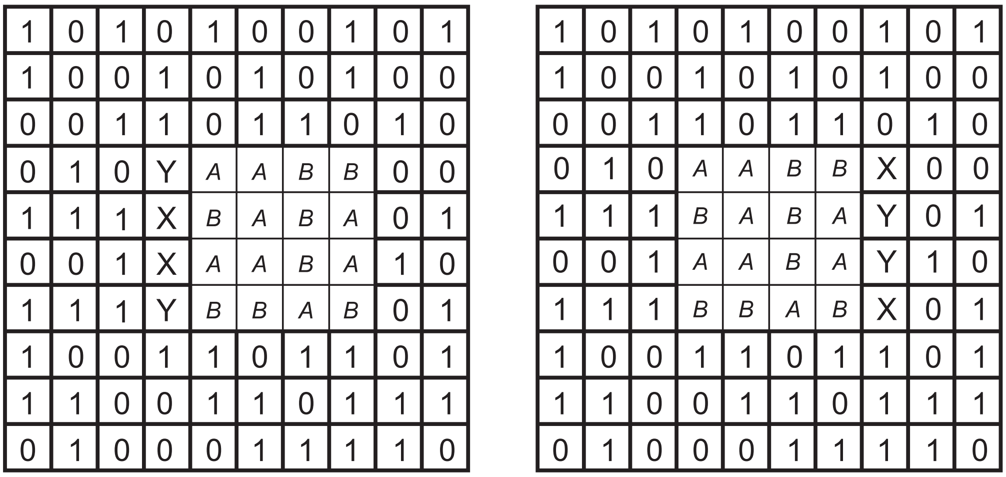
Figure 5: A random-dot stereogram from [31, Figure 2.4-1], reproduced with permission of Alcatel-Lucent/Bell Labs.

When the images in Figure 5 are viewed with a stereoscope, projecting the left image to the left eye and the right image to the right eye, the central square appears gradually as a discrete object suspended above the background.

Although this illustrates depth perception in stereoscopic vision—a subject of some interest in its own right—the main interest here is on how we see the central square as a discrete object. There is no such object in either of the two images individually. It exists purely in the *relationship* between the two images, and seeing it means matching one image with the other and unifying the parts which are the same.

This example shows that, although the matching and unification of patterns is a usefully simple idea, there are interesting subtleties and complexities that arise when two patterns are similar but not identical:

- Seeing the central object means finding a 'good' match between relevant pixels in the central area of the left and right images, and likewise for the background. Here, a good match is one that yields a relatively high level



1	0	1	0	1	0	0	1	0	1
1	0	0	1	0	1	0	1	0	0
0	0	1	1	0	1	1	0	1	0
0	1	0	Y	A	A	B	B	0	0
1	1	1	X	B	A	B	A	0	1
0	0	1	X	A	A	B	A	1	0
1	1	1	Y	B	B	A	B	0	1
1	0	0	1	1	0	1	1	0	1
1	1	0	0	1	1	0	1	1	1
0	1	0	0	0	1	1	1	1	0

1	0	1	0	1	0	0	1	0	1
1	0	0	1	0	1	0	1	0	0
0	0	1	1	0	1	1	0	1	0
0	1	0	A	A	B	B	X	0	0
1	1	1	B	A	B	A	Y	0	1
0	0	1	A	A	B	A	Y	1	0
1	1	1	B	B	A	B	X	0	1
1	0	0	1	1	0	1	1	0	1
1	1	0	0	1	1	0	1	1	1
0	1	0	0	0	1	1	1	1	0

Figure 6: Diagram to show the relationship between the left and right images in Figure 5. Reproduced from [31, Figure 2.4-3], with permission of Alcatel-Lucent/Bell Labs.

of information compression. Since there is normally an astronomically large number of alternative ways in which combinations of pixels in one image may be aligned with combinations of pixels in the other image, it is not normally feasible to search through all the possibilities exhaustively.

- As with many such problems in artificial intelligence, the best is the enemy of the good. Instead of looking for the perfect solution, we can do better by looking for solutions that are good enough for practical purposes. With this kind of problem, acceptably good solutions can often be found in a reasonable time with heuristic search, as in the SP system (Appendices C.1 and C.2): doing the search in stages and, at each stage, concentrating the search in the most promising areas and cutting out the rest, perhaps with backtracking or something equivalent to improve the robustness of the search. One such method for the analysis of random-dot stereograms has been described by Marr and Poggio [37].

7 Abstracting object concepts via motion

It seems likely that the kinds of processes that enable us to see a hidden object in a random-dot stereogram also apply to how we see discrete objects in the world. The contrast between the relatively stable configuration of features in an object such as a car, compared with the variety of its surroundings as it travels around, seems to be an important part of what leads us to conceptualise the object as an object [75, Section 5.2].

Any creature that depends on camouflage for protection—by blending with its background—must normally stay still. As soon as it moves relative to its surroundings, it is likely to stand out as a discrete object [75, Section 5.2] (see also Section 2.4.6).

The idea that information compression may provide a means of discovering 'natural' structures in the world—such as the many objects in our visual world—has been dubbed the 'DONSVIC' principle: *the discovery of natural structures via information compression* [74, Section 5.2]. Of course, the word 'natural' is not precise, but it has enough precision to be a meaningful name for the process of learning the kinds of concepts which are the bread-and-butter of our everyday thinking.

Similar principles may account for how young children come to understand that their first language (or languages) is composed of words (Section 10).

8 Adaptation in the eye of *Limulus* and run-length coding

Information compression may also be seen down in the works of vision. Figure 7 shows a recording from a single sensory cell (*ommatidium*) in the eye of a horseshoe crab (*Limulus polyphemus*), first when the background illumination is low, then when a light is switched on and kept on for a while, and later switched off—shown by the step function at the bottom of the figure.

As one might expect, the ommatidium fires at a relatively low rate of about 20 impulses per second even when the illumination is relatively low (shown at the left of the figure). When the light is switched on, the rate of firing increases sharply but instead of staying high while the light is on (as one might expect), it drops back almost immediately to the background rate. The rate of firing remains at that level until the light is switched off, at which point it drops sharply and then returns to the background level, a mirror image of what happened when the light was switched on.

For the main theme of this paper, a point of interest is that the positive spike when the light is switched on, and the negative spike when the light is switched off, have the effect of marking boundaries, first between dark and light, and later between light and dark. In effect, this is a form of run-length coding (Section 2.4.3). At the first boundary, the positive spike marks the fact of the light coming on. As long as the light stays on, there is no need for that information to be constantly repeated, so there is no need for the rate of firing to remain at a high level. Likewise, when the light is switched off, the negative spike marks the transition

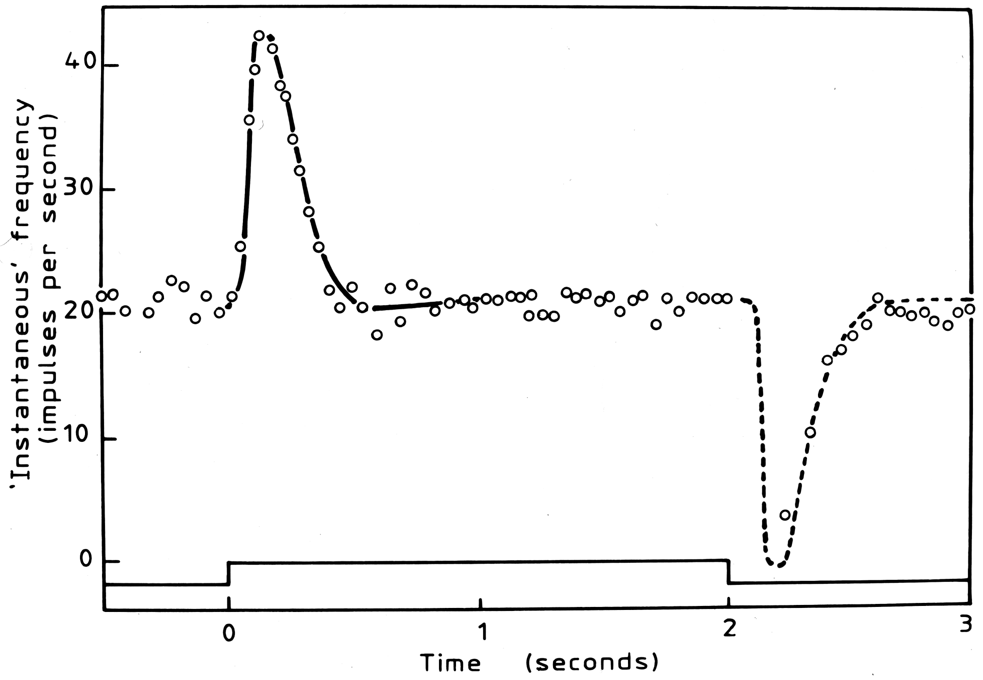


Figure 7: Variation in the rate of firing of a single ommatidium of the eye of a horseshoe crab in response to changing levels of illumination. Reproduced from [46, Figure 16], with permission from the Optical Society of America.

to darkness and, as before, there is no need for constant repetition of information about the new low level of illumination.[11]

Another point of interest is that this pattern of responding—adaptation to constant stimulation—can be explained via the action of inhibitory nerve fibres that bring the rate of firing back to the background rate when there is little or no variation in the sensory input [60]. Inhibitory mechanisms are widespread in the brain [56, p. 45] and it appears that, in general, their role is to reduce or eliminate redundancies in information ([76, Section 9], [44, Section 13.1]), in keeping with the main theme of this paper.

9 Other examples of adaptation

Adaptation is also evident at the level of conscious awareness. If, for example, a fan starts working nearby, we may notice the hum at first but then adapt to the sound and cease to be aware of it. But when the fan stops, we are likely to notice the new quietness at first but adapt again and stop noticing it.

Another example is the contrast between how we become aware if something or someone touches us but we are mostly unaware of how our clothes touch us in many places all day long. We are sensitive to something new and different and we are relatively insensitive to things that are repeated.

As with adaptation in the eye of *Limulus*, these other kinds of adaptation may be seen as examples of the run-length coding technique for compression of information.

10 Discovering the segmental structure of language

There is evidence that much of the segmental structure of language—words and phrases—may be discovered via ICMUP, as described in the following two subsections. To the extent that these mechanisms model aspects of HPLC, they provide evidence for BICMUP.

[11] It is recognised that this kind of adaptation in eyes is a likely reason for small eye movements when we are looking at something, including sudden small shifts in position (‘microsaccades’), drift in the direction of gaze, and tremor [38]. Without those movements, there would be an unvarying image on the retina so that, via adaptation, what we are looking at would soon disappear.

10.1 The word structure of natural language

As can be seen in Figure 8, people normally speak in 'ribbons' of sound, without gaps between words or other consistent markers of the boundaries between words. In the figure—the waveform for a recording of the spoken phrase "on our website"—it is not obvious where the word "on" ends and the word "our" begins, and likewise for the words "our" and "website". Just to confuse matters, there are three places within the word "website" that look as if they might be word boundaries.

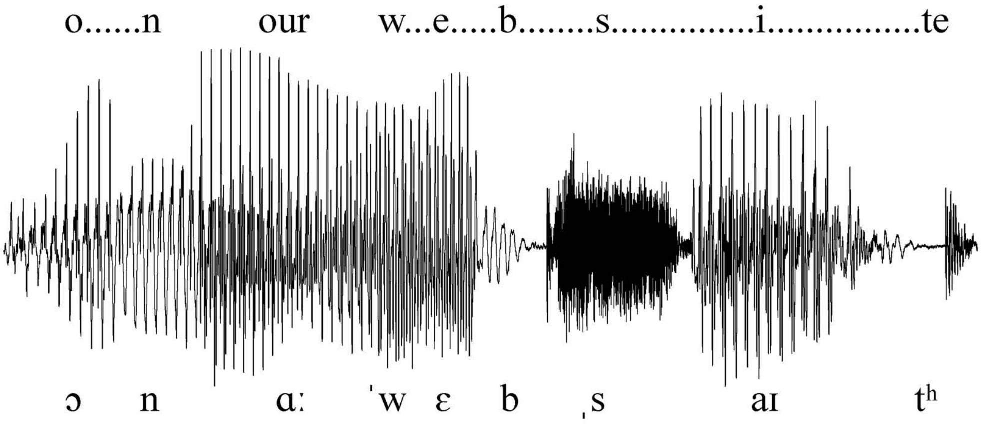


Figure 8: Waveform for the spoken phrase "On our website" with an alphabetic transcription above the waveform and a phonetic transcription below it. With thanks to Sidney Wood of SWPhonetics (swphonetics.com) for the figure and for permission to reproduce it.

Given that words are not clearly marked in the speech that young children hear, how do they get to know that language is composed of words? Learning to read could provide an answer but it appears that young children develop an understanding that language is composed of words well before the age when, normally, they are introduced to reading. Perhaps more to the point is that there are still, regrettably, many children throughout the world that are never introduced to reading but, in learning to talk and to understand speech, they inevitably develop a knowledge of the structure of language, including words.[12]

[12] It has been recognised for some time that skilled speakers of any language have an ability to create or recognise sentences that are grammatical but new to the world. Chomsky's well-known example of such a sentence is *Colorless green ideas sleep furiously.* [16, p. 15], which, when it was first published, was undoubtedly novel. This ability to create or recognise grammatical but novel sentences implies that knowledge of a language means knowledge of words as discrete entities that can form novel combinations.

In keeping with the main theme of this paper, ICMUP provides an answer. Computer model MK10 [66, 67, 70], which works largely via ICMUP, can reveal much of the word structure in an English-language text from which all spaces and punctuation has been removed [74, Section 5.2]. It true that there are added complications with speech but it seems likely that similar principles apply.

This discovery of word structure by the MK10 program, illustrated in Figure 9, is achieved without the aid of any kind of dictionary or other information about the structure of English. It is also achieved in 'unsupervised' mode, without the assistance of any kind of 'teacher', or data that is marked as 'wrong', or the grading of samples from simple to complex (*cf.* [25]). Statistical tests show that the correspondence between the computer-assigned word structure and the original (human) division into words is significantly better than chance.

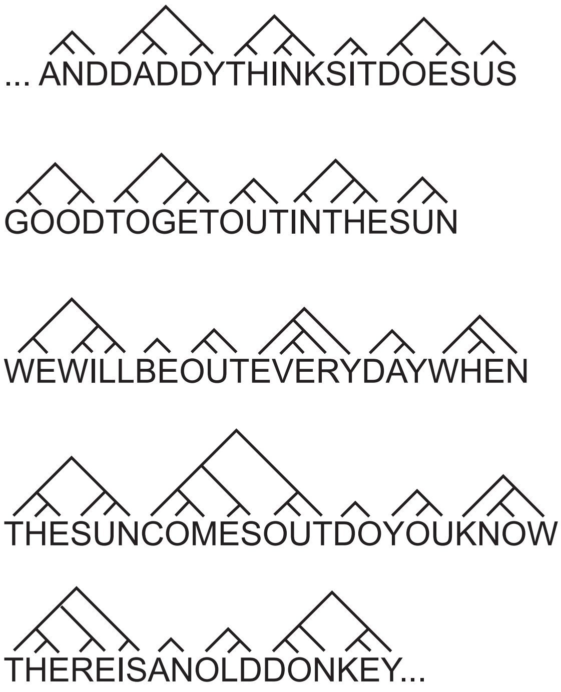


Figure 9: Part of a parsing created by program MK10 [67] from a 10,000 letter sample of English (book 8A of the Ladybird Reading Series) with all spaces and punctuation removed. The program derived this parsing from the sample alone, without any prior dictionary or other knowledge of the structure of English. Reproduced from Figure 7.3 in [70], with permission.

Two aspects of the MK10 model strengthen its position as a model of what children do in learning the segmental structure of language [67]: the growth in the lengths of words learned by the program corresponds quite well with the same measure for children; the pattern of changing numbers of new words that are learned by the program at different stages corresponds quite well with the equivalent pattern for children.

Discovering the word structure of language via ICMUP is another example of the DONSVIC principle, mentioned in Section 7—because words are the kinds of 'natural' structure which are the subject of the DONSVIC principle, and because ICMUP provides a key to how they may be discovered.

10.2 The phrase structure of natural language

Program MK10, featured in Section 10.1, does quite a good job at discovering the phrase structure of unsegmented text in which each word has been replaced by a symbol representing the grammatical class of the word [68, 70]. An example is shown in Figure 10. As before, the program works without any prior knowledge of the structure of English and, apart from the initial assignment of word classes, it works in unsupervised mode without the assistance of any kind of 'teacher', or anything equivalent. As before, statistical tests show that the correspondence between computer-assigned and human-assigned structures is statistically significant.[13]

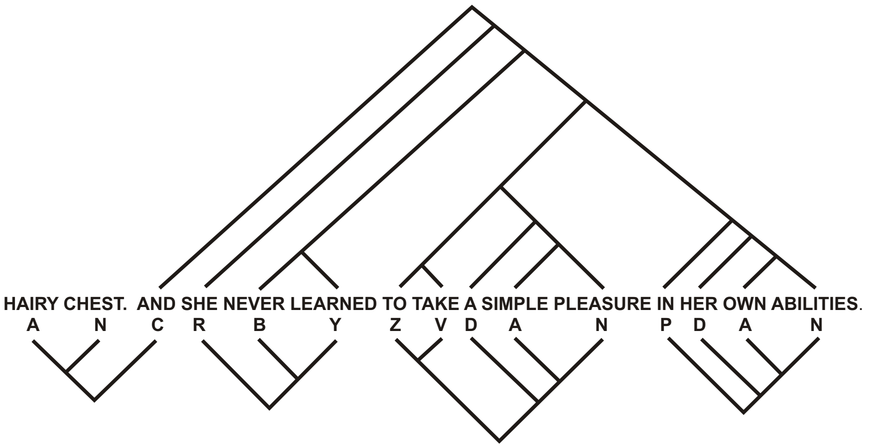

HAIRY CHEST. AND SHE NEVER LEARNED TO TAKE A SIMPLE PLEASURE IN HER OWN ABILITIES.
A N C R B Y Z V D A N P D A N

Figure 10: One sentence from a 7600 word sample from the book *Jerusalem the Golden* (by Margaret Drabble) showing (above the text) a surface structure analysis, and (below the text) the parsing developed by program MK10 at a late stage of processing [68]. This figure is reproduced by kind permission of Kingston Press Services Ltd.

Since ICMUP is central in the workings of the MK10 computer model, this

[13] Thanks to Dr. Isabel Forbes, a person qualified in theoretical linguistics, for the assignment of grammatical class symbols to words in the given text, and for phrase-structure analyses of the text.

result suggests that ICMUP may have a role to play, not merely in discovering the phrase structure of language, but more generally in discovering the grammatical structure of language (next).

11 Grammatical inference

Picking up the last point from the previous section, it seems likely that learning the grammar of a language may also be understood in terms of ICMUP. Evidence in support of that expectation comes from research with two programs designed for grammatical inference:

- Program SNPR, which was developed from program MK10, can discover plausible grammars from samples of English-like artificial languages [69, 70]. This includes the discovery of segmental structures, classes of structure, and abstract patterns. ICMUP is central in how the program works.
- Program SP71, one of the main products of the SP programme of research, achieves results at a similar level to that of SNPR. As before, ICMUP is central in how the program works. With the solution of some residual problems, outlined in [74, Section 3.3], there seems to be a real possibility that the SP system will be able to discover plausible grammars from samples of natural language. Also, it is anticipated that, with further development, the program may be applied to the learning of non-syntactic 'semantic' knowledge, and the learning of grammars in which syntax and semantics are integrated.

What was the point of developing SP71 when it does no better at grammatical inference than program SNPR? The reason is that the SNPR program, which was designed to build structures hierarchically, was not compatible with the new goal of the SP programme of research: to simplify and integrate observations and concepts across HLPC and related fields. What was needed was a new organising principle that would accommodate hierarchical structures and several other kinds of structure as well. It turns out that the SP-multiple-alignment concept is much more versatile than the hierarchical organising principle in the SNPR program, providing for the representation and processing of a variety of knowledge structures of which hierarchical structures is only one.

12 Generalisation, the correction of over- and under-generalisations, and 'dirty data'

Issues relating to generalisation in learning are best described with reference to the Venn diagram shown in Figure 11. It relates to the the unsupervised learning of a natural language but it appears that generalisation issues in other areas of learning are much the same.

The evidence to be described derives largely from the SNPR and SP computer models. Since both models are founded on ICMUP, evidence that they have human-like capabilities with generalisation and related phenomena may be seen as evidence in support of BICMUP.

In the figure, the smallest envelope shows the finite but large sample of 'utterances' (by adults and older children) from which a young child learns his or her native language (which we shall call **L**).[14] The next envelope shows the (infinite) set of utterances in **L**, and the largest envelope shows the (infinite) set of all possible utterances. 'Dirty data' are the many 'ungrammatical' utterances that children normally hear.

The child generalises 'correctly' when he or she infers **L** and nothing else from the finite sample he or she has heard, including dirty data. Anything that spills over into the outer envelope, like "mouses" as the plural of "mouse" or "buyed" as the past tense of "buy", is an over-generalisation, while failure to learn the whole of **L** represents under-generalisation.

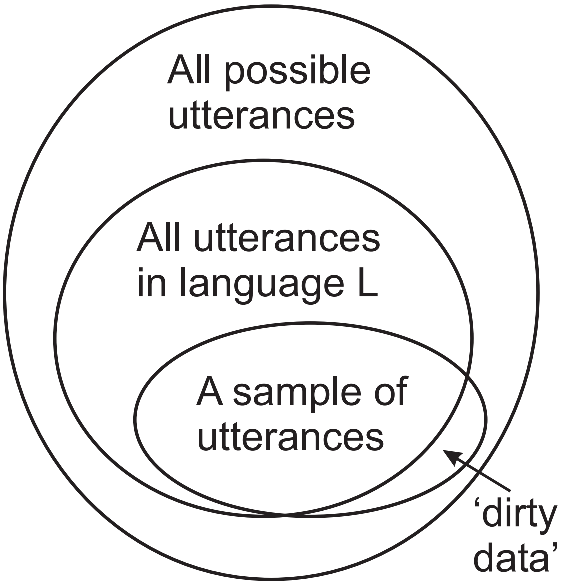


Figure 11: Categories of utterances involved in the learning of a first language, **L**. In ascending order size, they are: the finite sample of utterances from which a child learns; the (infinite) set of utterances in **L**; and the (infinite) set of all possible utterances. Adapted from Figure 7.1 in [70], with permission.

[14] To keep things simple in this discussion we shall assume that each child learns only one first language.

In connection with the foregoing summary of concepts relating to generalisation, there are three main problems:

- *Generalisation without over-generalisation.* How can we generalise our knowledge without over-generalisation, and this in the face of evidence that children can learn their first language or languages without the correction of errors by parents or teachers or anything equivalent?[15]

- *Generalisation without under-generalisation.* How can we generalise our knowledge without under-generalisation? As before, there is evidence that learning can be achieved without explicit correction of errors.

- *Dirty data.* How can we learn correct knowledge despite errors in the examples we hear. Again, it appears that this can be done without correction of errors.

These things are discussed quite fully in [72, Section 9.5.3] and [74, Section 5.3]. There is also relevant discussion in [77, Section V-H and XI-C].

In brief, information compression provides an answer to all three problems like this:

- For a given body of raw data, **I**, compress it thoroughly via unsupervised learning;

- The resulting compressed version of **I** may be split into two parts, a *grammar* and an *encoding* of **I** in terms of the grammar;

- Normally, the grammar generalises correctly without over- or under-generalisation, and errors in **I** are weeded out.

- The encoding may be discarded.

This scheme is admirably simple, but, so far, the evidence in support of it is only informal, deriving largely from informal experiments with English-like artificial languages with the SNPR computer model of language learning ([69], [70]) and the SP computer model [72, Section 9.5.3].

The weeding out of errors via this scheme may seem puzzling, but errors, by their nature, are rare. The grammar retains the repeating parts of **I** and discards

[15] Evidence comes chiefly from children who learned language without the possibility that anyone might correct their errors. Christy Brown was a cerebral-palsied child who not only lacked any ability to speak but whose bodily handicap was so severe that for much of his childhood he was unable to demonstrate that he had normal comprehension of speech and non-verbal forms of communication [8]. Hence, his learning of language must have been achieved without the possibility that anyone might correct errors in his spoken language.

the non-repeating parts including most of the errors (which are in the encoding). ‘Errors’ which are not rare acquire the status of ‘dialect’ and cease to be regarded as errors.

A problem with research in this area is that the identification of any over- or under-generalisations produced by the above scheme or any other model depends largely on human intuitions. But this is not so very different from the long-established practice in research on linguistics of using human judgements of grammaticality to establish what any given person knows about a particular language.

The problem of generalising our learning without over- or under-generalisation applies to the learning of a natural language and also to the learning of such things as visual images. It appears that the solution outlined here has distinct advantages compared with, for example, what appear to be largely *ad hoc* solutions that have been proposed for deep learning in artificial neural networks [77, Section V-H].

13 Perceptual constancies

It has long been recognised that our perceptions are governed by *constancies*:

- *Size constancy.* To a large extent, we judge the size of an object to be constant despite wide variations in the size of its image on the retina [21, pp. 40-41].

- *Lightness constancy.* We judge the lightness of an object to be constant despite wide variations in the intensity of its illumination [21, p. 376].

- *Colour constancy.* We judge the colour of an object to be constant despite wide variations in the colour of its illumination [21, p. 402].

These kinds of constancy, and others such as shape constancy and location constancy, may each be seen as a means of encoding information economically: it is simpler to remember that a particular person is “about my height” than many different judgements of size, depending on how far away that person is. In a similar way, it is simpler to remember that a particular object is “black” or “red” than all the complexity of how its lightness or its colour changes in different lighting conditions.

By filtering out variations due to viewing distance or the intensity or colour of incident light, we can facilitate ICMUP and thus, for example, in watching a football match, simplify the process of establishing that there is (normally) just one ball on the pitch and not many different balls depending on viewing distances, whether the ball is in a bright or shaded part of the pitch, and so on.

14 Mathematics

A discussion of mathematics may seem out of place in a paper about BICMUP but mathematics and computing are both products of human thinking that are designed to enhance the workings of the human mind. For that reason, in the spirit of George Boole's *An investigation of the laws of thought* [7], a consideration of their organisation and workings is relevant to the matter in hand.

In [78] it has been argued that mathematics may be seen as a set of techniques for the compression of information, and their application. In case this seems implausible:

- An equation like Albert Einstein's $E = mc^2$ may be seen as a very compressed representation of what may be a very large set of data points relating energy (E) and mass (m), with the speed of light (c) as a constant. Similar things may be said about such well-known equations as $s = (gt^2)/2$ (Newton's second law of motion), $a^2+b^2 = c^2$ (Pythagoras's equation), $PV = k$ (Boyle's law), and $F = q(E + v \times B)$ (the charged-particle equation).

- The second, third, and fourth of the variants of ICMUP outlined in Section 2.4.3 may be seen at work in mathematical notations. For example: multiplication as repeated addition may be seen as an example of run-length coding;

Owing to the close connections between logic and mathematics, and between computing and mathematics, it seems likely that similar principles apply in logic and in computing.

15 Evidence for bicmup via the SP system

Another strand of empirical evidence for BICMUP is via the SP computer model, which incorporates ICMUP within the SP-multiple-alignment construct. The model, with SP-multiple-alignment as its main component, demonstrates many features of HLPC. These are summarised quite fully in [79, Appendix B], and described in much more detail in [72, 74].

In summary, the strengths of the SP system in modelling aspects of HLPC are:

- *Versatility in the representation of knowledge.* The SP system has strengths in the representation of several different kinds of knowledge including: the syntax of natural languages; class-inclusion hierarchies (with or without cross classification); part-whole hierarchies; discrimination networks and trees; if-then rules; entity-relationship structures; relational tuples; and concepts in mathematics, logic, and computing, such as 'function', 'variable', 'value', 'set', and 'type definition'.

With the addition of two-dimensional SP patterns to the SP system, there is potential to represent such things as: photographs; diagrams; structures in three dimensions; and procedures that work in parallel.

- *Versatility in aspects of intelligence.* The SP system has strengths in several aspects of human-like intelligence including: unsupervised learning, the analysis and production of natural language; pattern recognition that is robust in the face of errors in data; pattern recognition at multiple levels of abstraction; computer vision; best-match and semantic kinds of information retrieval; several kinds of reasoning (next paragraph); planning; and problem solving.

  Strengths of the SP system in *reasoning* include: one-step 'deductive' reasoning; chains of reasoning; abductive reasoning; reasoning with probabilistic networks and trees; reasoning with 'rules'; nonmonotonic reasoning and reasoning with default values; Bayesian reasoning with 'explaining away'; causal reasoning; reasoning that is not supported by evidence; the already-mentioned inheritance of attributes in class hierarchies; and inheritance of contexts in part-whole hierarchies. There is also potential in the SP system for spatial reasoning and for what-if reasoning. Probabilities for inferences may be calculated in a straightforward manner.

- *Seamless integration of diverse kinds of knowledge and diverse aspects of intelligence, in any combination.* Because the SP system's versatility (in the representation of diverse kinds of knowledge and in diverse aspects of intelligence) flows from one relatively simple framework—SP-multiple-alignment—the system has clear potential for the seamless integration of diverse kinds of knowledge and diverse aspects of intelligence, in any combination. That kind of seamless integration appears to be essential in the modelling of HLPC.

16 Some apparent contradictions and how they may be resolved

The idea that ICMUP is fundamental in HLPC, and also in AI, mainstream computing, and mathematics, seems to be contradicted by:

- The productivity of the human brain and the ways in which computers and mathematics may be used to create redundant copies of information as well as to compress information;

- The fact that redundancy in information is often useful in both the storage and processing of information;

- A less direct challenge to BICMUP and the SP theory as a theory of HLPC is persuasive evidence, described by Gary Marcus [36], that in many respects, the human mind is a kluge, meaning "a clumsy or inelegant—yet surprisingly effective—solution to a problem" (*ibid.*, p 2).

- The fact that certain kinds of redundancy are difficult or impossible for people to detect and exploit.

These apparent contradictions and how they may be resolved are discussed in Appendix B.

17 Conclusion

This paper presents evidence for the idea, referred to as 'BICMUP', that much of the workings of brains and nervous systems may be understood as information compression via the matching and unification of patterns (ICMUP).

The paper is part of a programme of research developing the *SP theory of intelligence* and its realisation in the *SP computer model*—a theory which aims to simplify and integrate observations and concepts in human learning, perception, and cognition, and related areas.

Since ICMUP is a central part of the SP theory, evidence for BICMUP presented in this paper in Sections 4 to 14 inclusive (but excluding Section 15) strengthens empirical support for the SP theory, viewed as a theory of human learning, perception, and cognition.

Empirical evidence for the SP theory as a theory of human cognitive psychology—summarised in Section 15—provides evidence for BICMUP which is additional to that in Sections 4 to 14 inclusive.

Four possible objections to BICMUP and the SP theory are described, and how those objections may be answered.

Acknowledgements

I am grateful to anonymous reviewers for constructive and very helpful comments on an earlier version of this paper. Thanks also to Dr. Isabel Forbes for the assignment of grammatical class symbols to words in a given text, and for phrase-structure analyses, as described in Section 10.2.

A Barlow's change of view about the significance of information compression in mammalian learning, perception, and cognition, with comments

As noted in Section 3.1.1, Horace Barlow [6] argued that "... the [compression] idea was right in drawing attention to the importance of redundancy in sensory messages ... but it was wrong in emphasizing the main technical use for redundancy, which is compressive coding." (*ibid.* p. 242).

There are some valid points in what Barlow says in support of his new position but, as mentioned before, his overall conclusions appear to be wrong. His main arguments follow, with my comments after each one, flagged with 'JGW'.

1. "It is important to realize that redundancy is not something useless that can be stripped off and ignored. An animal must identify what is redundant in its sensory messages, for this can tell it about structure and statistical regularity in its environment that are important for its survival." [6, p. 243], and "It is ... knowledge and recognition of ... redundancy, not its reduction, that matters." [6, p. 244].

   JGW: It seems that the error here is to assume that compression of information means the complete elimination of redundant patterns. On the contrary, lossless compression of something like '`tabletabletabletabletable`' means retaining one instance of '`table`' with something to show the length of the sequence in the given context (Section 2.4.3).

   Knowledge of the frequency of occurrence of any pattern (in all its contexts) may serve in the calculation of absolute and relative probabilities ([72, Section 3.7], [74, Section 4.4]) and it can be the key to the correction of errors, as mentioned under point 2.

   In general, compression of information is entirely compatible with a knowledge of redundant patterns and what they can say about statistical regularities in a creature's environment that are important for its survival.

2. "Redundancy is mainly useful for error avoidance and correction" [6, p. 244]. This heading in [6] appears to be a relatively strong point in support of Barlow's new position, but he writes: "Since it is certainly true that sensory transducers and neural communication channels introduce noise, this is likely to be important in the brain, but the correction of such internally generated errors is a separate problem, and it will not be considered further here." [6, p. 244].

JGW: Redundancy can certainly be useful in the avoidance or correction of errors. But that does not invalidate BICMUP. As noted in Appendix B.2, the SP system, which is dedicated to the compression of information, will not work properly in such tasks as parsing, pattern recognition and grammatical inference, unless there are redundancies in its raw data. For that reason, it needs those redundancies in order to correct errors of omission, commission, and substitution, as described in [72, Section 6.2], [73, Section 2.2.2], and [74, Section 6.2].

In a similar way, the system can only work 'backwards' in decompression-by-compression (Appendix B.1.1) and in creating redundancy via information compression (Appendix B.1.2) if there is some redundancy that it can work on.

3. Following the remark that "This is the point on which my own opinion has changed most, partly in response to criticism, partly in response to new facts that have emerged." [6, p. 244], Barlow writes:

> "Originally both Attneave and I strongly emphasized the economy that could be achieved by recoding sensory messages to take advantage of their redundancy, but two points have become clear since those early days. First, anatomical evidence shows that there are very many more neurons at higher levels in the brain, suggesting that redundancy does not decrease, but actually increases. Second, the obvious forms of compressed, non-redundant, representation would not be at all suitable for the kinds of task that brains have to perform with the information represented; ..." [6, pp. 244–245].

and

> "I think one has to recognize that the information capacity of the higher representations is likely to be greater than that of the representation in the retina or optic nerve. If this is so, redundancy must increase, not decrease, because information cannot be created." [6, p. 245].

JGW: There seem to be two errors here:

- The likelihood that there are "very many more neurons at higher levels in the brain [than at the sensory levels]" and that "the information capacity of the higher representations is likely to be greater than that of the representation in the retina or optic nerve" does not in any way invalidate BICMUP.

It seems likely that many of the neurons at higher levels are concerned with the storage of one's accumulated knowledge over the period from one's birth to one's current age ([72, Chapter 11], [76, Section 4]). By contrast, neurons at the sensory level would be concerned only with the processing of sensory information at any one time.

Although knowledge in one's long-term memory stores is likely to be highly compressed and only a partial record of one's experiences, it is likely, for most of one's life except early childhood, to be very much larger than the sensory information one is processing at any one time. Hence, it should be no surprise to find many more neurons at higher levels than at the sensory level.

- For reasons given under point 4, there seem to be errors in the proposition that "the obvious forms of compressed, non-redundant, representation would not be at all suitable for the kinds of task that brains have to perform with the information represented."

4. Under the heading "Compressed representations are unsuitable for the brain", Barlow writes:

> "The typical result of a redundancy-reducing code would be to produce a distributed representation of the sensory input with a high activity ratio, in which many neurons are active simultaneously, and with high and nearly equal frequencies. It can be shown that, for one of the operations that is most essential in order to perform brain-like tasks, such high activity-ratio distributed representations are not only inconvenient, but also grossly inefficient from a statistical viewpoint ..." [6, p. 245].

JGW: With regard to these points:

- It is not clear why Barlow should assume that compressed representations are unsuitable for the brain, or that a redundancy-reducing code would, typically, produce a distributed representation as he describes. The SP system is dedicated to the creation of non-distributed compressed representations and, in [76], it is argued that such representations can be mapped on to plausible structures of neurons and their inter-connections that are quite similar to Donald Hebb's [26] concept of a 'cell assembly'.

- With regard to efficiency:
  - It is true that deep learning in artificial neural networks [51], with their distributed representations, are often hungry for computing resources. But otherwise they are quite successful with certain kinds of task, and there appears to be scope for increasing their efficiencies [15].
  - The SP system demonstrates that the compressed localist representations in the system are efficient and effective in a variety of kinds of task, as described in [72] and [74].

B Some apparent contradictions of BICMUP and the SP theory, and how they may be resolved

The apparent contradictions of BICMUP and the SP theory as a theory of HLPC that were mentioned in Section 16 are discussed in the following four subsections, with suggested answers to those apparent contradictions.

B.1 The creation of redundancy via information compression: 'decompression by compression'

The idea that information may be decompressed by compressing information—'decompression by compression'—seems paradoxical at first sight. Examples described here may help to show why the paradox is more apparent than real.

B.1.1 A simple example of 'decompression by compression'

In the retrieval of compressed information, the chunking-with-codes idea outlined in Section 2.4.3 provides a simple example of decompression by compression:

- *Compression of information.* If, for example, a document contains many instances of the expression "Treaty on the Functioning of the European Union" we may shorten it by giving that expression a relatively short name or code like 'TFEU' and then replacing all but one instances of the long expression with its shorter code. This achieves compression of information because, in effect, multiple instances of "Treaty on the Functioning of the European Union" have been reduced to one via matching and unification.

- *Retrieval of compressed information.* We can reverse the process and thus decompress the document by searching for instances of 'TFEU' and replacing each one with "Treaty on the Functioning of the European Union". But to achieve that result, the search pattern, 'TFEU', needs to be matched and unified with each instance of 'TFEU' in the document. And that process of matching and unification is itself a process of compressing information. Hence, decompression of information has been achieved via information compression.

B.1.2 Creating redundancy via information compression

With a computer, it is very easy to create information containing large amounts of redundancy and to do it by a process which may itself be seen to entail the compression of information.

We can, for example, make a 'call' to the function defined in Figure 12, using the pattern '`oranges_and_lemons(100)`'. The effect of that call is to print out a highly redundant sequence containing 100 copies of the expression 'Oranges and lemons, Say the bells of St. Clement's; '.

```
void oranges_and_lemons(int x)
{
     printf("Oranges and lemons, Say the bells of St. Clement's; ");
     if (x > 1) oranges_and_lemons(x - 1) ;
}.
```

Figure 12: A simple recursive function showing how, via computing, it is possible to create repeated (redundant) copies of 'Oranges and lemons, Say the bells of St. Clement's; '.

Taking things step by step, this works as follows:

1. The pattern '`oranges_and_lemons(100)`' is matched with the pattern '`void oranges_and_lemons(int x)`' in the first line of the function.

2. The two instances of '`oranges_and_lemons`' are unified and the value 100 is assigned to the variable x. The assignment may also be understood in terms of the matching and unification of patterns but the details would be a distraction from the main point here.

3. The instruction '`printf("Oranges and lemons, Say the bells of St. Clement's; ");`' in the function has the effect of printing out '`Oranges and lemons, Say the bells of St. Clement's; `'.

4. Then if $x > 1$, the instruction '`oranges_and_lemons(x - 1)`' has the effect of calling the function again but this time with 99 as the value of x (because of the instruction $x - 1$ in the pattern '`oranges_and_lemons(x - 1)`', meaning that 1 is to be subtracted from the current value of x).

5. Much as with the first call to the function (item 1, above), the pattern '`oranges_and_lemons(99)`' is matched with the pattern '`void oranges_and_lemons(int x)`' in the first line of the function.

6. Much as before, the two instances of '`oranges_and_lemons`' are unified and the value 98 is assigned to the variable x.

7. This cycle continues until the value of x is 0.

Where does compression of information come in? It happens mainly when one copy of '`oranges_and_lemons`' is matched and unified with another copy so that, in effect, two copies are reduced to one.

There is more about recursion at the end of Appendix B.1.3, next.

B.1.3 Decompression by compression and the creation of redundancy via the SP system

How the SP system may achieve decompression by compression is described in [72, Section 3.8] and [74, Section 4.5]. It would not be appropriate to reproduce that description in this paper, and in any case the details are not needed here. Three points are relevant:

- Decompression of a body of information **I**, may be achieved by a process which is *exactly* the same as the process that achieved the original compression of **I**: there is no modification to the program of any kind.

- All that is needed to achieve decompression is to ensure that there is some residual redundancy in the compressed version of **I**, so that the program has something to work on, as noted at the end of point 2 in Appendix A.

- The SP computer model is entirely devoted to compression of information, without any special provision for decompression of information.

Those three things establish that it is indeed possible to achieve decompression by compression, meaning that, in that idea, there is really no paradox or contradiction.

With regard to the creation of redundancy via recursion discussed in Appendix B.1.2, readers may find it useful to examine examples of recursion with the SP system, described in [72, Sections 4.3.2.1 and 5.3], [75, Section 3.3], and [76, Section

7]. In all these examples, recursion is driven by a process which is unambiguously devoted to the compression of information. And this is true in examples like [72, Figure 4.4a] where recursion has the effect of creating redundancy in information.

B.2 Redundancy is often useful in the storage and processing of information

The fact that redundancy—repetition of information—is often useful in both the storage and processing of information is the second apparent contradiction to BICMUP and the SP theory as a theory of HLPC. Here are some examples:

- With any kind of database, it is normal practice to maintain one or more backup copies as a safeguard against catastrophic loss of the data. Each backup copy represents redundancy in the system.

- With information on the internet, it is common practice to maintain two or more 'mirror' copies in different places to minimise transmission times and to spread processing loads across two or more sites, thus reducing the chance of overload at any one site. Again, each mirror copy represents redundancy in the system.

- Redundancies in natural language can be a very useful aid to the comprehension of speech in noisy conditions.

- It is normal practice to add redundancies to electronic messages, in the form of additional bits of information together with checksums, and also by repeating the transmission of any part of a message that has become corrupted. These things help to safeguard messages against accidental errors caused by such things as birds flying across transmission beams, or electronic noise in the system, and so on.

Since similar principle apply in biological systems (Section 4), these examples appear to argue against BICMUP. However, in both artificial and natural systems, uses of redundancy of the kind just described may co-exist with ICMUP. For example: "... it is entirely possible for a database to be designed to minimise internal redundancies and, at the same time, for redundancies to be used in backup copies or mirror copies of the database ... Paradoxical as it may sound, knowledge can be compressed and redundant at the same time." [72, Section 2.3.7].

As we have seen at the end of point 2 in Appendix A and in Appendix B.1.3, the SP system, which is dedicated to the compression of information, will not work properly with totally random information containing no redundancy. It needs redundancy in its 'New' data in order to achieve such things as the parsing of natural language, pattern recognition, and grammatical inference. Also, for the correction of errors in any incoming batch of New SP-patterns, it needs a repository of Old patterns that represent patterns of redundancy in a previously-processed body of New information.

B.3 The human mind as a kluge

As mentioned in Section 16, Gary Marcus has described persuasive evidence that, in many respects, the human mind is a kluge. To illustrate the point, here is a sample of what Marcus says:

> "Our memory is both spectacular and a constant source of disappointment: we recognize photos from our high school year-books decades later—yet find it impossible to remember what we had for breakfast yesterday. Our memory is also prone to distortion, conflation, and simple failure. We can know a word but not be able to remember it when we need it ... or we can learn something valuable ... and promptly forget it. The average high school student spends four years memorising dates, names, and places, drill after drill, and yet a significant number of teenagers can't even identify the *century* in which World War I took place." [36, p. 18, emphasis as in the original].

Clearly, human memory is, in some respects, much less effective than a computer disk drive or even a book. And it seems likely that at least part of the reason for this and other shortcomings of the human mind is that "Evolution [by natural selection] tends to work with what is already in place, making modifications rather than starting from scratch." and "piling new systems on top of old ones" [36, p. 12].

Superficially, this and other evidence in [36] seems to undermine the idea that there is some grand unifying principle—such as information compression via SP-multiple-alignment—that governs the organisation and workings of the human mind.

Perhaps, as Marvin Minsky suggested, "each [human] mind is made of many smaller processes" called *agents* each one of which "can only do some simple thing that needs no mind or thought at all. Yet when we join these agents in societies—in certain very special ways—this leads to true intelligence." [41, p. 17].

In answer to these points:

- The evidence that Marcus presents is persuasive: it is difficult to deny that, in certain respects, the human mind is a kluge. And evolution by natural selection provides a plausible explanation for anomalies and inconsistencies in the workings of the human mind.

- But those conclusions are entirely compatible with BICMUP and the SP theory as a theory of mind. As Marcus says:

  > "I don't mean to chuck the baby along with its bath—or even to suggest that kluges outnumber more beneficial adaptations. The biologist Leslie Orgel once wrote that 'Mother Nature is smarter than you are,' and most of the time it is." [36, p. 16].

  although Marcus warns that in comparisons between artificial systems and natural ones, nature does not always come out on top.

In general it seems that, despite the evidence for kluges in the human mind, there can be powerful organising principles too. Since BICMUP and the SP theory are well supported by evidence, they are likely to provide useful insights into the nature of human intelligence, alongside an understanding that there are likely to be kluge-related anomalies and inconsistencies too.

Minsky's counsel of despair—"The power of intelligence stems from our vast diversity, not from any single, perfect principle." [41, p. 308]—is probably too strong. It is likely that there is at least one unifying principle for human-level intelligence, and there may be more. And it is likely that, with people, any such principle or principles operates alongside the somewhat haphazard influences of evolution by natural selection.

B.4 Some kinds of redundancy are difficult or impossible for people to detect and exploit

There is no doubt that people are imperfect in their abilities to detect and exploit redundancy. For example:

> "... a grid in which pixels encoded the binary expansion of π would, of course, have a very simple description, but this structure would not be identified by the perceptual system; the grid would, instead, appear completely unstructured." [9, p. 578].

At first sight, this shortfall in our abilities seems to undermine the idea that ICMUP is a unifying principle in the workings of brains and nervous systems. But that idea does not imply that brains and nervous systems are perfect compressors of information. Indeed, it appears that with all but the smallest or most regular bodies of information, it is necessary to use heuristic techniques for compression of the information and that these cannot guarantee that the best possible result has been found (Appendices C.1 and C.2, [77, Appendix I-E.4]).

C Outline of the sp theory of intelligence

As mentioned in Section 2.2, the *SP theory of intelligence* and its realisation in the *SP computer model* is a unique attempt to simplify and integrate observations and concepts across artificial intelligence, mainstream computing, mathematics, and human learning, perception, and cognition, with information compression as a unifying theme.

As noted in that section, the name 'SP' is short for *Simplicity* and *Power*, because compression of any given body of information, **I**, may be seen as a process of reducing informational 'redundancy' in **I** and thus increasing its 'simplicity', whilst retaining as much as possible of its non-redundant expressive 'power'.

The theory, the computer model, and some applications, are described most fully in [72] and more briefly in [74]. Details of other publications, most with download links, may be found on www.cognitionresearch.org/sp.htm.

The SP theory is conceived as a brain-like system as shown schematically in Figure 13. The system receives *New* information via its senses and stores some or all of it in compressed form as *Old* information.[16]

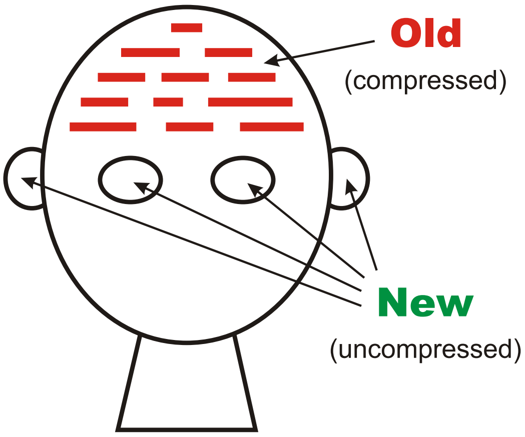


Figure 13: Schematic representation of the SP system from an 'input' perspective. Reproduced with permission from Figure 1 in [74].

All kinds of knowledge or information in the SP system are represented with arrays of atomic *symbols* in one or two dimensions called *SP-patterns*.[17] At present

[16] To avoid unnecessary confusion, the SP computer model is dedicated exclusively to lossless compression of information. But with such a model it always possible to discard information if that proves necessary in the modelling of human psychology or in applications of the projected industrial-strength *SP machine*, based on the SP theory and the SP computer model [44].

[17] Up until recently, the concept of of an array of atomic symbols in the SP system has been

the SP computer model works only with one-dimensional SP-patterns but it is envisaged that, at some stage, it will be generalised to work with two-dimensional SP-patterns.

C.1 SP-multiple-alignment

A central part of the SP system is the powerful concept of *SP-multiple-alignment*, outlined here. The concept is described more fully in [72, Sections 3.4 and 3.5] and [74, Section 4]

The concept of SP-multiple-alignment in the SP system is derived from the concept of 'multiple alignment' in bioinformatics.[18,19] That latter concept means an arrangement of two or more DNA sequences or sequences of amino-acid residues so that, by judicious 'stretching' of sequences in a computer, matching symbols are aligned—as illustrated in Figure 14. A 'good' multiple alignment is one with a relatively high value for some metric related to the number of symbols that have been brought into line.

```
  G G A     G     C A G G G A G G A     T G     G   G G A
  | | |     |     | | | | | | | | |     | |     |   | | |
  G G | G   G C C C A G G G A G G A     | G G C G   G G A
  | | |     | | | | | | | | | | | |     | |     |   | | |
A | G A C T G C C C A G G G | G G | G C T G     G A | G A
  | | |           | | | | | | | | |   |   |     |   | | |
  G G A A         | A G G G A G G A   | A G     G   G G A
  | |   |         | | | | | | | |     |   |     |   | | |
  G G C A         C A G G G A G G     C   G     G   G G A
```

Figure 14: A 'good' multiple alignment amongst five DNA sequences. Reproduced with permission from Figure 3.1 in [72].

For a given set of sequences, finding or creating 'good' multiple alignments amongst the many possible 'bad' ones is normally a complex process—normally too complex to be solved by exhaustive search. For that reason, bioinformatics programs for finding good multiple alignments use heuristic methods, building multiple alignments in stages and discarding low-scoring multiple alignments at the end of each stage. With such methods it is not normally possible to guarantee that

referred to as 'pattern'. But now the term *SP-pattern* has been introduced to mark the distinction between arrays of symbols in the SP system from the more general concept of pattern, which may include both whole SP-patterns and portions of such SP-patterns.

[18] See "Multiple sequence alignment", *Wikipedia*, http://bit.ly/2nArijz, retrieved 2017-10-11.

[19] Up until recently, the concept of 'multiple alignment' in the SP system has been referred to as 'multiple alignment'. But now the term *SP-multiple-alignment* has been introduced to mark the distinction and the important differences between the concept of multiple alignment as it has been developed in the SP system from the concept of multiple alignment in bioinformatics.

the best possible multiple alignment has been found, but it is normally possible to find multiple alignments that are good enough for practical purposes.

The main difference between the concept of SP-multiple-alignment in the SP system and the concept of multiple alignment in bioinformatics is that:

- With an SP-multiple-alignment, one of the SP-patterns (sometimes more than one) is *New* information from the system's environment (see Figure 13), and the remaining SP-patterns are *Old* information, meaning information that has been previously stored (also shown in Figure 13).

  In the creation of SP-multiple-alignments, the aim is to build ones that, in each case, allow the New SP-pattern (or SP-patterns) to be encoded economically in terms of the Old SP-patterns in the given SP-multiple-alignment. In each case, there is an implicit merging or unification of SP-patterns or parts of SP-patterns that match each other, as described in [72, Section 3.5] and [74, Section 4.1].

- With multiple alignments in bioinformatics, all the sequences (of DNA bases or amino-acid residues etc) have the same status. In general, the aim in creating multiple alignments in bioinformatics is to develop ones with relatively large values for some kind of measure related to the numbers of symbols that are brought into line.

In the SP-multiple-alignment shown in Figure 15, one New SP-pattern is shown in row 0 and Old SP-patterns, drawn from a repository of Old SP-patterns, are shown in rows 1 to 9, one SP-pattern per row. By convention, the New SP-pattern(S) is always shown in row 0 and the Old SP-patterns are shown in the other rows, one SP-pattern per row.

In this example, the New SP-pattern is a sentence and the SP-patterns in rows 1 to 9 represent grammatical structures including words. The overall effect of the SP-multiple-alignment is to 'parse' or analyse the sentence into its constituent parts and sub-parts, with each part marked with a category like '`NP`' (meaning 'noun phrase'), '`N`' (meaning 'noun'), '`VP`' (meaning 'verb phrase'), and so on.

With SP-multiple-alignments in the SP system, as with multiple alignments in bioinformatics, the process of finding 'good' SP-multiple-alignments is too complex for exhaustive search, so it is normally necessary to use heuristic methods—which means that it is normally not possible to guarantee that the best possible SP-multiple-alignment has been found, but it is normally possible to find SP-multiple-alignments that are good enough for practical purposes.

At the heart of SP-multiple-alignment is a process for finding good full and partial matches between SP-patterns, described quite fully in [72, Appendix A].

```
0              f o r t u n e                      f a v o u r     s             t h e        b r a v e               0
               | | | | | | |                      | | | | | |     |             | | |        | | | | |
1              | | | | | | |                 Vr 6 f a v o u r #Vr |             | | |        | | | | |               1
               | | | | | | |                 |                 |  |             | | |        | | | | |
2              | | | | | | |             V 7 Vr               #Vr s #V          | | |        | | | | |               2
               | | | | | | |             |                          |           | | |        | | | | |
3              | | | | | | |        VP 3 V                          #V NP       | | |        | | | | |    #NP #VP    3
               | | | | | | |        |                                  |        | | |        | | | | |     |   |
4          N 4 f o r t u n e #N     |                                  |        | | |        | | | | |     |   |     4
           |                 |      |                                  |        | | |        | | | | |     |   |
5     NP 2 N                 #N #NP |                                  |        | | |        | | | | |     |   |     5
      |                          |  |                                  |        | | |        | | | | |     |   |
6 S 0 NP                        #NP VP                                 |        | | |        | | | | |     |  #VP #S 6
                                                                       |        | | |        | | | | |     |
7                                                                      |        | | |    N 5 b r a v e #N  |         7
                                                                       |        | | |    |             |   |
8                                                                      NP 1 D   | | | #D N             #N #NP        8
                                                                            |   | | | |
9                                                                           D 8 t h e #D                             9
```

Figure 15: The best SP-multiple-alignment produced by the SP computer model with a New SP-pattern representing a sentence to be parsed and a repository of user-supplied Old SP-patterns representing grammatical categories, including words. Reproduced with permission from Figure 2 in [76].

C.2 Unsupervised learning in the SP system

Unsupervised learning in the SP system is outlined in [79, Appendix A.4] and described more fully in [74, Section 5] and [72, Chapter 9]. As with the building of SP-multiple-alignments, unsupervised learning in the SP system uses heuristic techniques so that, normally, it is not possible to guarantee that the best possible result has been found.

C.3 Strengths and potential of the SP system

The strengths and potential of the SP system, in the representation of diverse kinds of knowledge, in diverse aspects of intelligence, and in the seamless integration of diverse kinds of knowledge and diverse aspects of intelligence in any combination, is summarised in [79, Appendix B] with pointers to where fuller information may be found.

C.4 SP-neural

The SP system, including the concept of SP-multiple-alignment with ICMUP, suggests how aspects of intelligence may be realised in a 'neural' version of the SP theory, *SP-neural* expressed in terms of neurons and their interconnections [76].

An important point here is that SP-neural is quite different from the kinds of 'artificial neural network' that are popular in computer science, including those that provide the basis for 'deep learning' [51]. And learning in the SP system is fundamentally different from deep learning in artificial neural networks.

It is relevant to mention here that [77, Section V] describes thirteen problems with deep learning in artificial neural networks and how, with the SP system, those problems may be overcome.

The SP system also provides a comprehensive solution to a fourteenth problem with deep learning—"catastrophic forgetting"—meaning the way in which new learning in a deep learning system wipes out old memories. A solution has been proposed in [33] but it appears to be partial, and it is unlikely to be satisfactory in the long run.

C.5 Distinctive features and advantages of the SP system

Distinctive features and advantages of the SP system are described quite fully in [77] and outlined in [79, Appendix A.6].

C.6 Potential benefits and applications of the SP system

Potential benefits and applications of the SP system are summarised in [79, Appendix A.7] with pointers to relevant publications.

References

[1] L. Allison and C. S. Wallace. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *Journal of Molecular Evolution*, 39:418–430, 1994.

[2] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[3] H. B. Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In HMSO, editor, *The Mechanisation of Thought Processes*, pages 535–559. Her Majesty's Stationery Office, London, 1959.

[4] H. B. Barlow. Trigger features, adaptation and economy of impulses. In K. N. Leibovic, editor, *Information Processes in the Nervous System*, pages 209–230. Springer, New York, 1969.

[5] H. B. Barlow. Intelligence, guesswork, language. *Nature*, 304:207–209, 1983.

[6] H. B. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253, 2001.

[7] G. Boole. *An Investigation of the Laws of Thought.* Walton and Maberly, London, Kindle edition, 1854.

[8] C. Brown. *My Left Foot.* Vintage Digital, London, Kindle edition, 2014. First published in 1954.

[9] N. Chater. Reconciling simplicity and likelihood principles in perceptual organisation. *Psychological Review*, 103(3):566–581, 1996.

[10] N. Chater. The search for simplicity: a fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, 52 A(2):273–302, 1999.

[11] N. Chater and M. Oaksford. *The Probabilistic Mind: Prospects for Bayesian Cognitive Science.* Oxford University Press, Oxford, 2008.

[12] N. Chater and P. Vitányi. Simplicity: a unifying principle in cognitive science? *TRENDS in Cognitive Sciences*, 7(1):19–22, 2003.

[13] N. Chater and P. Vitányi. 'Ideal learning' of natural language: positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3):135–163, 2007.

[14] M. Chekaf, N. Cowan, and F. Mathy. Chunk formation in immediate memory and how it relates to data compression. *Cognition*, 155:96–107, 2016.

[15] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman. Project adam: building an efficient and scalable deep learning training system. In *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 571–582. USENIX Association, 2014.

[16] N. Chomsky. *Syntactic Structures.* Mouton, The Hague, 1957.

[17] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, New York, 1991.

[18] D. C. Donderi. Visual complexity: a review. *Psychological Bulletin*, 132(1):73–97, 2006.

[19] R. Falk and C. Konold. Making sense of randomness: implicit encoding as a basis for judgment. *Psychological Review*, 104(2):301, 1997.

[20] J. Feldman. Minimization of boolean complexity in human concept learning. *Nature*, 407(6804):630–633, 2000.

[21] J. P. Frisby and J. V. Stone. *Seeing: The Computational Approach to Biological Vision.* The MIT Press, London, England, 2010.

[22] N. Gauvrit, H. Singmann, F. Soler-Toscano, and H. H. Zenil. Algorithmic complexity for psychology: a user-friendly implementation of the coding theorem method. *Behavior Research Methods*, 48(1):314–329, 2016.

[23] N. Gauvrit, F. Soler-Toscano, and H. Zenil. Natural scene statistics mediate the perception of image complexity. *Visual Cognition*, 22(8):1084–1091, 2014.

[24] B. Goertzel. Cogprime: an integrative architecture for embodied artificial general intelligence. Technical report, The Open Cognition Project, 2012. PDF: bit.ly/1Zn0qfF, 2012-10-02.

[25] M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[26] D. O. Hebb. *The Organization of Behaviour.* John Wiley & Sons, New York, 1949.

[27] J. Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of algorithmic complexity. In *Proceedings of the International Symposium of Engineering of Intelligent Systems (EIS '98)*, pages 146–163, 1998.

[28] A. S. Hsu, N. Chater, and P. Vitáyi. Language learning from positive evidence, reconsidered: a simplicity-based approach. *Topics in Cognitive Science*, 5:35–55, 2013.

[29] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2005. ISBN 3-540-22139-5, www.hutter1.net/ai/uaibook.htm.

[30] M. Hutter. One decade of universal artificial intelligence. In P. Wang and B. Goertzel, editors, *Theoretical Foundations of Artificial General Intelligence*, volume 4, pages 67–88. Springer, Heidelberg, 2012.

[31] B. Julesz. *Foundations of Cyclopean Perception.* Chicago University Press, Chicago, 1971.

[32] S. Kirby, T. Griffiths, and K. Smith. Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114, 2014.

[33] J. Kirkpatrick. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017.

[34] B. Lemaire, V. Robinet, and S. Portrat. Compression mechanisms in working memory. *Mathématiques et Sciences Humaines*, 199(3):71–84, 2012.

[35] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 3rd edition, 2014.

[36] G. Marcus. *Kluge: the Hapharzard Construction of the Human Mind*. Faber and Faber, London, paperback edition, 2008. ISBN: 978-0-571-23652-7.

[37] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B*, 204(1156):301–328, 1979.

[38] S. Martinez-Conde, J. Otero-Millan, and S. L. Macknik. The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, 14:83–96, 2013.

[39] F. Mathy and J. Feldman. What's magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, 122(3):346–362, 2012.

[40] G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

[41] M. Minsky, editor. *The Society of Mind*. Simon & Schuster, New York, 1986.

[42] A. Newell. You can't play 20 questions with nature and win: projective comments on the papers in this symposium. In W. G. Chase, editor, *Visual Information Processing*, pages 283–308. Academic Press, New York, 1973.

[43] A. Newell, editor. *Unified Theories of Cognition*. Harvard University Press, Cambridge, Mass., 1990.

[44] V. Palade and J. G. Wolff. Development of a new machine for artificial intelligence. Technical report, CognitionResearch.org, 2017. Submitted for publication. bit.ly/2tWb88M, arXiv:1707.0061.

[45] E. M. Pothos and N. Chater. A simplicity principle in unsupervised human categorization. *Cognitive Science*, 26:303–343, 2002.

[46] F. Ratliff, H. K. Hartline, and W. H. Miller. Spatial and temporal aspects of retinal inhibitory interaction. *Journal of the Optical Society of America*, 53:110–120, 1963.

[47] J. Rissanen. Modelling by the shortest data description. *Automatica*, 14(5):465–471, 1978.

[48] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society B*, 49(3):223–239, 1987.

[49] V. Robinet, B. Lemaire, and M. B. Gordon. Mdlchunker: a mdl-based cognitive codel of inductive learning. *Cognitive Science*, 35:1352–1389, 2011.

[50] K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann, Amsterdam, 2012.

[51] J. Schmidhuber. Deep learning in neural networks: an overview. *Neural Networks*, 61:85–117, 2015.

[52] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[53] R. R. Sokal and P. H. A. Sneath, editors. *Numerical Taxonomy: the Principles and Practice of Numerical Classification*. W. H. Freeman, San Francisco, 1973.

[54] R. J. Solomonoff. A formal theory of inductive inference. Parts I and II. *Information and Control*, 7:1–22 and 224–254, 1964.

[55] R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.

[56] L. R. Squire, D. Berg, F. E. Bloom, S. du Lac, A. Ghosh, and N. C. Spitzer, editors. *Fundamental Neuroscience*. Elsevier, Amsterdam, fourth edition, 2013.

[57] M. Tamariz and S. Kirby. Culture: copying, compression and conventionality. *Cognitive Science*, 39(1):171–183, 2015.

[58] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A monte-carlo aixi approximation. *Journal of Artificial Intelligence Research*, 40(1):95–142, 2011.

[59] P. Vitányi and N. Chater. Identification of probabilities. *Journal of Mathematical Psychology*, 76(Part A):13–24, 2017.

[60] G. von Békésy. *Sensory Inhibition*. Princeton University Press, Princeton, NJ, 1967.

[61] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–195, 1968.

[62] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society B*, 49(3):240–252, 1987.

[63] S. Watamabe. Information-theoretical aspects of inductive and deductive inference. *IBM Journal of Research and Development*, 4:208–231, 1960.

[64] S. Watanabe, editor. *Frontiers of Pattern Recognition*. Academic Press, New York, 1972.

[65] S. Watanabe. Pattern recognition as information compression. In *Frontiers of Pattern Recognition* [64].

[66] J. G. Wolff. An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology*, 66:79–90, 1975. See `www.cognitionresearch.org/lang_learn.html#wolff_1975`.

[67] J. G. Wolff. The discovery of segments in natural language. *British Journal of Psychology*, 68:97–106, 1977. bit.ly/Yg3qQb.

[68] J. G. Wolff. Language acquisition and the discovery of phrase structure. *Language & Speech*, 23:255–269, 1980. See `www.cognitionresearch.org/lang_learn.html#wolff_1980`.

[69] J. G. Wolff. Language acquisition, data compression and generalization. *Language & Communication*, 2:57–89, 1982. doi.org/10.1016/0271-5309(82)90035-0, bit.ly/Zq0zAl.

[70] J. G. Wolff. Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine, editors, *Categories and Processes in Language Acquisition*, pages 179–215. Lawrence Erlbaum, Hillsdale, NJ, 1988. bit.ly/ZIGjyc.

[71] J. G. Wolff. Computing, cognition and information compression. *AI Communications*, 6(2):107–127, 1993. bit.ly/XL359b.

[72] J. G. Wolff. *Unifying Computing and Cognition: the SP Theory and Its Applications*. CognitionResearch.org, Menai Bridge, 2006. ISBNs: 0-9550726-0-3 (ebook edition), 0-9550726-1-1 (print edition). Distributors, including Amazon.com, are detailed on bit.ly/WmB1rs.

[73] J. G. Wolff. Towards an intelligent database system founded on the SP theory of computing and cognition. *Data & Knowledge Engineering*, 60:596–624, 2007. bit.ly/1CUldR6, arXiv:cs/0311031.

[74] J. G. Wolff. The SP theory of intelligence: an overview. *Information*, 4(3):283–341, 2013. bit.ly/1NOMJ6l, arXiv:1306.3888.

[75] J. G. Wolff. Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision. *SpringerPlus*, 3(1):552–570, 2014. bit.ly/2oIpZB6, arXiv:1303.2071.

[76] J. G. Wolff. Information compression, multiple alignment, and the representation and processing of knowledge in the brain. *Frontiers in Psychology*, 7:1584, 2016. bit.ly/2esmYyt, arXiv:1604.05535.

[77] J. G. Wolff. The SP theory of intelligence: its distinctive features and advantages. *IEEE Access*, 4:216–246, 2016. bit.ly/2qgq5QF, arXiv:1508.04087.

[78] J. G. Wolff. On the "mysterious" effectiveness of mathematics in science. Technical report, CognitionResearch.org, 2017. Submitted for publication. bit.ly/2otrHD0, viXra:1706.0004, hal-01534622.

[79] J. G. Wolff. Software engineering and the SP theory of intelligence. Technical report, CognitionResearch.org, 2017. Submitted for publication. bit.ly/2w99Wzq, arXiv:1708.06665.

[80] J. G. Wolff. Strengths and potential of the sp theory of intelligence in general, human-like artificial intelligence. Technical report, CognitionResearch.org, 2017.

[81] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[82] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.