Cognitive Bootstrapping: A Survey of Bootstrap Mechanisms for Emergent Cognition

David Windridge
School of Electronics & Physical Sciences
University of Surrey,
Guildford, Surrey.

Abstract

We propose 'Cognitive Bootstrapping' as a blanket term for all instances the process by which the perceptual apparatus of an autonomous agent is conceptually extended and experimentally validated. Bootstrap techniques are necessary to transcend the paradox inherent in validating perceptual categorisations via the objects of perception. Perception may thus become self-founding only within certain crucial a priori limits required to maintain referentiality and provide a validation criterion for the proposed perceptual updates. We hence survey the subject areas in which this mechanism occurs, ultimately advocating a hierarchically open-ended perception/action approach to artificial cognition in order to objectively ground perceptual updating.
Cognitive Bootstrapping:
A Survey of Bootstrap Mechanisms for Emergent Cognition in Natural and Artificial Agents

DAVID WINDRIDGE
University of Surrey

November, 2005
Contents

I INTRODUCTION

0.1 The Relevance of Bootstrapping to Artificial Cognition ................. 7
  0.1.1 The Central Conceptual Difficulty of Artificial Cognition ........ 7
  0.1.2 The Bootstrap Solution ........................................ 7
  0.1.3 The Underlying Argument of the Survey ....................... 10
0.2 Organisation .................................................................. 13
0.3 A Note On Terminology .................................................. 14

II PHILOSOPHICAL PERSPECTIVES ON COGNITIVE BOOTSTRAPPING .... 16

1 Cognition and its Relation to the Objective World ...................... 16
  1.1 The Kantian View of Cognition ..................................... 16
    1.1.1 Transcendental Unity of Apperception ...................... 18
    1.1.2 Objecthood ..................................................... 19
    1.1.3 Kant and Cognitive Bootstrapping ......................... 21

2 Hermeneutics: The Hermeneutic Circle and Cognitive bootstrapping 23
  2.1 Textual Hermeneutics ............................................... 23
  2.2 Non-Textual Hermeneutics .......................................... 24
  2.3 The Necessity of Embodiment ...................................... 26
  2.4 Ontology-Free Views of Embodied Cognition and the Possibility of Cognitive Bootstrapping ........................................ 27
    2.4.1 Quine ......................................................... 27
    2.4.2 Wittgenstein .................................................. 28
  2.5 Implications for the Achievability of Artificial Cognition ........ 29
    2.5.1 Theoretical Implications: The Very Possibility Of Artificial Cognition .............................................. 29
    2.5.2 Practical Implications: Real World Implementation of the Hermeneutic Circle ........................................... 30
  2.6 Summary of Division .................................................. 33

III COGNITIVE SCIENCE PERSPECTIVES ON COGNITIVE BOOTSTRAPPING 34
3 General Cognitive Science 34
3.1 Introduction .................................................. 34
3.2 Approaches to Cognitive Science .............................. 34
   3.2.1 Cognitive Bootstrapping in Symbolic Accounts of Human Cognition 35
   3.2.2 Cognitive Bootstrapping in Connectionist Approaches to Cognitive Science .................................. 36
   3.2.3 Convergence of Symbolic and Connectionist Schools ............... 39
3.3 The Significance of the Embodied Mind Within Cognitive Science .... 41

4 Cognitive Linguistic Context for Cognitive Bootstrapping 43
4.1 Section Introduction: The Origins of Symbolic Representation in Cognitive Agents ........................................ 43
4.2 Origins of Language in the Symbolic Interchange of Percepts .......... 44
4.3 Language Evolution in Embodied Artificial Cognitive Agents ........ 46
   4.3.1 General Ontology Learning Mechanisms ...................... 46
   4.3.2 Spontaneous Language Formation ............................ 50
4.4 Summary of Division ............................................ 53

IV MATHEMATICAL PERSPECTIVES ON COGNITIVE BOOTSTRAPPING 54

5 Foundational Context for Cognitive Bootstrapping 54
5.1 Formalism Verses Platonism .................................. 54
5.2 The Middle Ground of Constructivism .......................... 54
5.3 Cognitive Bootstrapping and Constructivism ................... 55

6 Statistical Context for Cognitive Bootstrapping 57
6.1 Bayesian Updating and Cognitive Bootstrapping .................. 57
   6.1.1 Bayesian Exploration ........................................ 59
   6.1.2 Dirichlet Processes: A Framework for Non-Parametric Bayesian Updating .................................. 60
6.2 Algorithmic-Information Theoretic Frameworks for Fully Unconstrained Bayesian Updating ......................... 62
   6.2.1 The Bayesian Prior Problem ................................. 63
   6.2.2 Bayes Theory and the Minimum Description Length Principle ... 63
   6.2.3 A Universal Prior for Bayesian Updating .................. 64
   6.2.4 Practical Implementation of Algorithmic Bayesian Updating .... 65
V COMPUTATIONAL PERSPECTIVES ON COGNITIVE BOOTSTRAPPING 67

6.3 Summary of Division ........................................ 66

7 Cognitive Bootstrapping Methods in the Field of Artificial Intelligence 67
7.1 Robotics ...................................................... 68
7.1.1 Open-Ended Cognitive Bootstrapping in Autonomous Robots ... 68
7.1.2 Constrained Cognitive Bootstrapping in Autonomous Robots ... 74
7.2 Non-Physically Embodied Cognitive Bootstrapping Mechanisms In Computer Science .................................................. 76
7.3 Summary of Division ........................................ 77

VI DISCUSSION AND CONCLUSIONS 79

8 Discussion of Philosophical Issues 82
8.1 The Necessity of Bootstrapped Cognition and Embodied Agency ...... 82
8.2 Implications for Artificial Cognition ................................ 84
8.3 Self-Objectivising Cognitive Models ................................ 85
8.3.1 Relative verses Absolute Agency ............................ 86

9 Discussion of Cognitive Science Issues 87
9.1 Overcoming the Problem of Perceptual Reference in Open-ended Cognition 87
9.1.1 First Stage: Construction of the Bootstrap Percept Hypothesis .. 87
9.1.2 Second Stage: Updating the Bootstrap Percept Hypothesis ....... 88
9.2 Symbol Grounding and Intentionality; Hierarchically ................ 89

10 Conclusions 90
10.1 Summary of Requirements for Cognitive Bootstrapping in Artificial Cognitive Models .............................................. 90
10.1.1 A Priori Requirements .................................... 91
10.1.2 A Posteriori Requirements .................................. 92
10.2 Formal List of Requirements for Cognitive Bootstrapping ........... 92
Part I

0.1 The Relevance of Bootstrapping to Artificial Cognition

0.1.1 The Central Conceptual Difficulty of Artificial Cognition

It has been a well established principle, dating from the conception of computing, that a human-equivalent intelligence is not to be straightforwardly designed by human means; to do so would to involve, as a first step, knowing exactly one’s own mental structures, and hence embracing the fundamental foundational paradox of Bertrand Russell. This paradox, as significant for the development of mathematics as it was for computing science, arises when the permissibility of sets containing as members other sets (in particular the permissibility of set self-membership) is granted as logically acceptable\(^1\). In so far as mental representations may be modelled by set-theoretic constructs, Russell’s paradox has thus to be addressed within the domain of artificial intelligence.

From this paradox, and others derivable of it (for instance, those surrounding putative algorithmic solutions to the Halting problem \([137]\), or the possibility of complete, finite axiomatizations of natural number theory \([46]\)), we have become used to questioning the possibility of a finite system (such as the human brain) encompassing a complete self-representation within itself, particularly one that is demonstrably such (within that system). Of course, partial or temporally-retrograde self-models are permissible, so that it is hence possible for a human-being to use a linguistic token ‘I’ meaningfully and accurately, or, on a computationally level, to build mobile robots capable of building accurate models of their position in space, if not of their full internal state-space. However, absolute and immediate self-models would appear to be ruled-out completely (see for instance \([13]\) for a discussion of the limits to self-observation under finite, Markovian and infinite state-space assumptions, and \([12]\) under quantum-physical assumptions).

It would thus appear that the mechanics of human cognition, if indeed such a notion is permissible, could not be knowingly expressed in a finite and formally complete manner by any human being, and could not therefore be implemented via conventional methods of computational engineering, wherein a set of a set of functional goals are first specified before being algorithmically implemented (for an extended discussion of this notion cf \([74]\)). This is, in effect, to transpose the negative conclusion of the Hilbert programme (the attempt, in the 1920s to construct, in advance, a formal axiomatisation of all mathematics) from a mathematical context to that of cognitive science, where the laws of cognition are hence the quantity that is incapable of a provably - ie knowingly- complete analytic formulation.

0.1.2 The Bootstrap Solution

Clearly, though, human-level cognition exists in a replicable fashion throughout the world, reproduced through well understood biological and chemical mechanisms. The

---
\(^1\)In the strictest sense, paradox only occurs when considering sets of non-self-membering sets; self-membership alone causes only problems of infinite regress. However, formal completeness requires that the existence of the latter construct implies the existence of the former construction, and hence the paradox cannot easily be evaded; we have at the very least to deal with explicit infinities in any such domain.
idea of creating artificial cognitive entities is not therefore ruled out in advance by the laws of physics. The difficulty lies only with the algorithmic specification of the problem, which is necessarily not accessible to human thought. As early as the 1950s [49] it had been realised that an embodied, machine-learning approach provides the only obvious resolution of this problem, wherein a machine is constructed capable of reprogramming its own, originally human-specified low-level perceptual code (constituting a partial human self-model), in terms of learnt higher-level conceptual percepts deriving from the lower level code. However, the combined mechanical and computational requirements of such an endeavour were such that it is only in the latter part of the 20th century that this approach to the creation of artificial cognitive systems has been attempted; traditional attempts at artificial cognition have rested on well understood, rigorously defined mathematical foundations, which, if not (via the above arguments) constitutive of complete theories of human cognition, have had the advantage of being both knowably applicable and effective. The emerging research field of self-determining artificial cognitive systems then represents the goal of replicating human-like cognition at both a hardware and software level, with the dynamic link between the two being as essential to this endeavour as the individual components. Charting the development of these systems will form a central component of this survey, which will also need, to an extent, to assess historical developments in the field of classical fixed-form artificial cognitive agency in order to appropriately contextualise the open-ended alternative.

This paradigm for artificial cognitive systems, the reprogramming of lower-level perceptual code in terms of learnt higher-level conceptual percepts might itself appear paradoxical; indeed it is a version of the classical computational 'bootstrap' whereby (for instance) the compiler for a new computational language is written within that same language. The key requirement that prevents this notion from being as paradoxical as the legendary attempts of Baron Munchhausen to fly by 'pulling himself up by his own bootstraps' (from whence we get the expression), is that the core terms of the new language must rest on a pre-existing (a priori) computational-linguistic foundation, even if the final form of the compiled language retains no explicit trace of its origins. We would hence like to propose the term Cognitive Bootstrapping to describe this 'self-programming' form of solution to the central difficulty of artificial cognition, particularly in so far as it refers to the updating of the perceptual machinery of cognition, and not simply the updating of an environmental model in terms of fixed perceptual categorisations. A central difficult of cognitive bootstrapping is thus that environmental and perceptual model refinement must be carried out simultaneously despite their mutual interdependency.

In fact, ideas of equivalent functional form to that underlying the above definition of cognitive bootstrapping have a long historical pedigree, arising in many different forms within many different subject areas. For instance, it occurs in Statistics (where it is formalised as a particular subset of Bayesian updating [eg [5]], Philosophy (in the form of Ricoeur's notion of the hermeneutic arc [114]), Cognitive science (where it occurs in both the symbolist and connectionist theories of mind), Cybernetics (where it occurs within mechanisms such as simultaneous localisation and mapping), Linguistics (where it occurs in Pinker's model of language acquisition [105]) and, finally, Pure-mathematics (where it can be found at the core of foundational theories of mathematics, such as Chaitin's work on empirical axiom selection [21], that imposes what Hofstadter [55] might describe as a 'tangled hierarchy' of axiom/theorem relations). The central,
paradoxical notion of cognitive self-assessment and modification implicit in cognitive bootstrapping might even be seen in more diverse areas, such as literature. For the requirements of this survey, we therefore propose that the term cognitive bootstrapping further serves as an umbrella term for all manifestations of this idea in the various fields.

Cognitive bootstrapping is hence much more than merely unsupervised cognitive learning; this would not involve overcoming the chicken-and-egg paradox inherent in an agent with unlimited capacity for forming novel percept categories with which to view the world, nonetheless being able to perceive whether these categories are representative or not of the world. (Unsupervised cognitive learning would not require the learner to interpret sensory data in terms of its classification of that same data, and to evaluate that interpretation in the light of new data gained via that interpretation; unsupervised cognitive learning has an externally-imposed success criterion, and can treat the sensory data as being, in a sense, it own ground truth - cognitive bootstrapping, on the other hand, must treat its own current representations as the nearest approximation to a ground truth, since these are all that are relevant to it as an embodied and active agent, and conduct cognitive updating in terms of these ungrounded representations).

Utilising bootstrapping to overcome this paradox requires that we first obtain an initial set of low-level percept categories that we know are 'correct' (which is to say, representative) a priori, and only then progressing to higher-levels of the perceptual hierarchy. This initial category set, we shall argue, is the set of Kantian synthetic a priori cognitive categories that provide a framework in which Popperian falsification of percept category hypotheses can be meaningfully formulated, with high-level percepts constructed as a hypothesised conjunction of low-level percepts. Validation of novel percept hypotheses is achieved by 'projecting' them back into the environment as percept-action linkages predicated on the assumption of their representativity. These conjectures hence accumulate in complexity from the initial bootstrap hypothesis, with the bootstrapping agent alternating between hypothesising and exploration phases in the same manner that an autonomous mobile robot progressively builds-up and refines its environment map by utilising the partial maps to further explore the areas of the objective environment in which the map is not yet adequately determined.

The concept of cognitive bootstrapping is thus analogous to the mechanism of semantic learning that we employ as infants, in which we must first obtain a sufficient (bootstrap) sub-set of words and word-meanings in order to be able to formulate questions concerning meaning of new words, and thus expand our vocabulary indefinitely. For an investigator who observed only the later querying phase of language learning, and inferred that words and word definitions were only learned through being defined in terms of query-responses constructed from other words, the presence of language in any individual would appear paradoxical (since the language learner would have had to self-referentially ask for the

\footnote{For instance, in one interpretation of Shakespeare’s The Tempest, the key protagonist Prospero also acts, at a meta-level, as the play’s author, determining all of his own and the audiences perceptions of events within the play. Only by writing himself into the play as an agent, acted upon by the plays events, rather than simply being their omniscient initiatior and interpreter, is he finally, in the last act, able to view his narrative perceptions for what they are; merely limited representations of an objective world. By testing his perceptions as an embodied agent, Prospero is thus able to overcome the paradox involved in determining the applicability of his own self-determined perceptual categories and to enter reality (to the point of addressing the audience directly), thereby becoming both author of and actor within the world (Prospero is hence commonly identified with Shakespeare, who had both of these roles).}
meaning of query-related words like 'ask' and 'meaning'). The resolution of the paradox relies, of course, on the fact that the initial bootstrap word-set has meanings grounded both in perceptual correlations (such as onomatopoeism) and on principles of reference and hypothesis-formation that are 'hard-wired' into the human brain; in effect, an a priori category of word meanings.

0.1.3 The Underlying Argument of the Survey

It is thus evident that, as well as begin a survey of the various contexts in which the cognitive bootstrap mechanism occurs, this article must also serve as an argument for its application to artificial cognition, as well as an argument for the particular form that it must take within the field. Exposition of this argument will occur throughout the whole of the article, as the requirements imposed upon artificial cognition by the differing disciplines become apparent.

In the broadest terms, however, this argument can be specified from both the philosophical and cognitive science perspectives, and may be summarised briefly as follows.

1. The Philosophical Argument. As we have indicated, the philosophical argument of the survey revolves around a central paradox: how can a cognitive agent capable of changing its perceptual categories (that is, its way of seeing the world) ever validate one particular set of perceptual categories over another? The concept of validation involves, at the very least, the perception of the inadequacy of one perceived entity in relation to another; but how can perceptual categories ever be objects of perception?

This is not soluble in terms of either the Cartesian or Classical Empiricist schools of philosophy, since the first claims cognitive agents cannot absolutely validate the existence of anything beyond their own percepts, and the second does not recognise the possibility of the perceptual mediation of the objective world (objects present themselves as they are 'in themselves' directly to cognition).

Kant, however, provides an alternative conceptual framework, asserting that an object is, by a priori cognitive necessity, that which exists outside of an agent's percept domain, being rather that to which percepts refer. Percepts hence serve to mediate between agent and object, being crucial to their distinction as separate entities. Objects are thus never perceived by cognitive agents as they are in themselves (being required to conform to the a priori requirements of perception): however, nor are they simply reducible to percepts. Instead, object concepts are accessible to cognition as ordering concepts imposed on intuitions (singular, low-level sensory percepts), rather than being themselves singular percepts: they are (in Kant's jargon) synthetic unities.

These object concepts are thus of an inherently hypothetical nature, existing beyond the certainty of the current sensory impressions, serving instead as proposed linkages between those impressions. Were they confirmable to cognition, however, these object concepts would themselves have the status of singular, higher-level percepts, which could in turn serve as the basis for further synthetically unified object-concepts. Object hypotheses thus link lower-level percepts together in a conjectural unity (such as, for instance, non-simultaneous perspectives on a discrete three-dimensional object such as a building-brick), and have the capability, if sufficiently validated, to form hierarchies with the object hypothesis then referring only to its highest level (for instance, particular
classes of building made from bricks, if the certainty of the conception of 'brick' is sufficiently well established observationally).

However, percepts are (particularly in the philosophical school of phenomenology), linked together via actions. Actions thus inherently have the capability to test object hypotheses, falsifying those that do not have perception-action results equivalent to those proposed by the object hypothesis. Thus, in the above example, we would need to walk around a building in order to establish whether the entity present to the senses conforms to our class conception. We hence also see that the earlier semantic bootstrapping example can be made to conform to this framework by imposing the equivalences: perception=word; object hypothesis=word definition; hypothesis-testing action=querying. Actions thus test the consistency of observed percept linkages with respect to the underlying a priori object conjecture that serves to give unity to our perceptions (complementing the a priori unity of the perceiver).

In the Kantian/phenomenological framework it is thus possible, via this consistency condition (which is again, a priori), for an autonomous cognitive agent to update and validate its own perceptual categories (which is to say, engage in cognitive bootstrapping), but only by proceeding via a bottom-up perception-action hierarchy built on the assumption of the a priori referentiality of the lowest level of the agent’s percept domain with respect to the transcendental object, and the a priori consistency and relevance of the lowest-level of the agent’s motor space (so that, for instance, an autonomous robot is not meaningfully free to query the topology of its motor-space independently of its perceptual-space).

2. The Cognitive Science Argument  From the perspective of cognitive science it is possibly to give an entirely different argument for the form that cognitive bootstrapping must take from that above, but to arrive at exactly the same conclusion. This time, the argument is framed in terms of the problem of symbol grounding, which is particularly apparent in the construction of autonomous artificial agents. An autonomous cognitive agent is, by definition, one capable of adapting to its in environment in behavioural and representational terms that go beyond those implied by its initial set of 'bootstrap' symbolic assumptions, in order to find representations more suited to the particular environment in which the agent finds itself. Doing so necessitates the use of mechanisms of generalisation, inference and decision making in order to modify the initial perceptual symbol set in the light of novel forms of sensory data.

Any representation that is capable of abstract generalisation is implicitly governed by the laws of predicate logic. As such, the generalised entities must observe strictly formalised laws of interrelationship, and consequently, in abstracting the symbol set away from the original set of innate percept-behavioural pairings, are apt to become detached from any intrinsic meaning in relation to the agent’s environment. A related difficulty, known as the frame problem [82], also arises in such generalised formal domains; it is by no mean clear which particular set of logical consequences (given the infinite number of possibilities) that the generalised reasoning system should concern itself with.

There is hence a problem of symbol relevance and 'grounding' unless additional mechanisms can be put in place to form a bridge between the formal requirements of logical inference as applied to visual symbols, and the further constraint of the relevance of this symbol set to the agent within the context of both its goals and the intrinsic nature
of the environment in which these goals are to be fulfilled. In terms of the philosophy of cognition, this necessitates a move from a Quinean [110] to a Wittgensteinian [150] frame of reference, in which symbol meaning is intrinsically contextual, and environment-dependent, rather than being a matter of arbitrary ontological assumption.

For cognitive agents in the animal kingdom the grounding of symbols is guaranteed by the mechanism of Darwinian natural selection; representations that do not meaningfully and efficiently represent the survival prerogatives of the agent in the context of its environment increase the likelihood of its extinction and genetic removal from the heredity of future generations [87]. This mechanism, however, is not readily available to artificial cognitive agents other than in the context of self-replicating agents within a simulated environment (see eg [127] for an overview of this sub-field).

For artificial cognitive agents embodied within the real world (that is to say, robots), the form that this symbol grounding framework must take is, by an increasing consensus (eg [78], [40], [47], [63]), one of hierarchical stages of abstraction that proceed from the 'bottom-up'. At the lowest level is thus the immediate relationship between percept and action; a change in what is perceived is primarily brought about by actions in the agent's motor-space. This hence limits visual learning to what is immediately relevant to the agent, and significantly reduces the quantity of data from which the agent must construct its symbol domain by virtue of the many-to-one mapping that exists between the pre-symbolic visual space and the intrinsic motor space [76]. Thus, for example, a hypothetical mobile robot engaged in simultaneous location and mapping (SLAM) (eg [135]) might build up a stationary stochastic model of any environmental changes that occur when not engaged in any direct motor activity, but switches to a Markovian transitional model when engaged in motor activity, thereby forming a sequence of 'key-frame' transitions driven by its motor impulses.

The first level of abstraction in the hierarchy thus represents a generalisation of the immediate, pre-symbolic percept-action relation into the symbol domain. There are many approaches to achieving this primary generalisation, for instance: unsupervised clustering [76], invariant subspace factoring [48], constructive solid geometry schematics [22]. Progressive levels of abstraction can be added by similar means, or they might instead involve higher levels of inferential machinery, for instance first order logical induction for rule inference, if explicitly ascending the Chomsky hierarchy [24].

At some level of abstraction, critically, is the concept of objects, characterised by their persistence with respect to the agent's actions. Representations above this level are then characterised by their object-centric, rather than agent-centric descriptions (so we hence move from a percept-action space into a region in which descriptions with formal equivalents to English terms such as 'on', 'under', etc, can form part of the environment description).

We hence end up with a set of high-level, abstracted symbol generalisations which are nevertheless grounded in the percept-action space by virtue of the intermediate hierarchical levels. We might thus, for instance, envisage a tennis-playing robot that has the segmentation of the ball from the background at its lowest representative level, leading into a series of ascending representations that cumulate in the formal logical rules of the game of tennis at the most abstract level of representation.

Such a hierarchical structure has the further characteristic that higher-level action imperatives (such as, in our example, 'serving the ball') act to reinforce learning at the

12
lower-levels (by providing additional tennis-ball segmentation statistics at the lowest-level, for instance). When this learning is intentional (such as when a mobile robot utilises an inferred high-level environmental representation as a default perceptual hypothesis for interpreting further environment data in order that it can explore poorly mapped areas, and thereby drive the low-level data collection process by which further perceptual and environmental model refinement is accomplished), the system is then inherently one of cognitive-bootstrapping.

It is thus apparent that classical A.I. approaches to artificial cognitive were of limited success in that they attempted to build a high-level environmental description directly from the percept space before going on to consider agent actions within this model, rather than allowing this representation to evolve at a higher hierarchical level of the percept-action relation [16]. Representative priorities were thus specified in advance by the system-builder and not by the agent, meaning that autonomous agency had to build its goals and higher-level representations in terms of the assumed representational modes, with all the redundancy that this implied. Furthermore, novel modes of representation were frequently ruled out in advance by this pre-specification of scene-description.

It is hence possible to provide an argument in favour of a hierarchical, bottom-up, perception-action-based approach to the construction of artificial cognition from a purely cognitive-science-centric perspective that implicitly incorporates cognitive bootstrapping, but which is independent of (and therefore complementary to) the preceding philosophical argument.

0.2 Organisation

As regards the purely surveying aspect of this article, we shall find it worthwhile to investigate all of the associated subject areas indicated above in some detail, and at an individual level, in order to summarise relevant bootstrapping research in the field of cognition, artificial and otherwise. This is particularly so as each of the areas addresses very much the same underlying conception, but from within the conceptual language of their respective underlying fields. It may thereby prove possible to refine our means of addressing the concept of cognitive bootstrapping further; indeed this process of arriving at the central notion via a series of partial formulations is itself reminiscent of cognitive bootstrapping. The research process itself can thus perhaps be considered our most universal exemplar of the cognitive bootstrap technique.

Hence, the four principle parts into which this survey is divided concern the Philosophical, the Cognitive, the Mathematical and the Computational Science contexts of cognitive bootstrapping theory. Later sections of each of the part-divisions are thus concerned with examining previous and ongoing efforts to construct self-founding artificial cognitive systems within the differing conceptual frameworks. It will be necessary, where appropriate, to outline artificial cognitive research that falls into the 'fixed-function', non-bootstrapping mould, insofar as it sheds light on our central concern of self-founding open-ended cognitive architectures.

The overall document structure, Philosophical: Cognitive: Mathematical: Computational, is hence one of increasing specificity, the earlier parts outlining the idea of cognitive bootstrapping in the most abstract and general terms, with the later parts dealing with actual issues of implementation. This structure is also, as indicated, imposed out
internally within each chapter, giving a consistent overall continuity of format.

Inevitably, boundaries between these subject-divisions are porous to an extent, with each division including some elements of the others. We shall therefore, in adopting the chosen four-fold partitioning, concern ourselves principally with the core ideological progressions of the respective schools, treating any particular areas of overlap within their single most representative sphere. The survey is thus far from exhaustive, but seeks rather to give a representation of the scope of cognitive bootstrapping techniques across the range of academic disciplines in so far as they have a bearing on the problem of artificial cognition.

The conclusion section will draw together these strands, summarising the a priori and empirical necessities of cognitive bootstrapping, and propose a checklist of elements necessary to a cognitive architecture in order to qualify as one of cognitive bootstrapping.

0.3 A Note On Terminology

In order to convey as much as possible to the non-specialist reader, we shall use the term 'percept' to refer to singular and basic sensory impressions, whether qualia or more complex constructs. The critical defining aspect of a percept is sensory immediacy and perceived singularity. The term 'percept categories' shall then denote groupings of percepts into collective entities in such a manner that it can further serve as a singular percept (by the above definition). Percept categories are thus perceptual particulars existing at hierarchical levels above that of the innate sensory impressions. Hence, a set of individual percepts consisting of filled rectilinear facets might be collectively resolved as the percept category 'brick', which in turn might be collectively resolved (as individual percepts) into the percept category 'building'. This may also happen on an abstract level, whereby the collectivity of percepts is logical, rather than set theoretical (as in the above examples). Hence, the percept category 'objects which appear above other objects' collects all particular instances of the ordered percept pairs \( x, y \) where there exists a relation \( \text{height}\{x\} > \text{height}\{y\} \). The terminology is thus general enough to allow for mechanisms of unsupervised percept clustering as well as logical inference in determining new percept categories.

However, it should be noted that some of the argument of the following article exists within an explicitly Kantian framework, where the use of some of these terms is rather different. Specifically, in Kantian terms, concepts actively ensure that perceptions conform to the categories, so that object representations can exist for cognitive agents in general relationships (spatial, temporal etc). A category is thus the most basic form of concept, examples being quantity, unity, plurality, totality, relation, and causality. Intuitions are the passive form of object representation, consisting of the particulars of sensations. Concepts are thus generalised groupings of intuitions. Of central importance is the concept of a transcendental object, which is an a priori necessity for cognition. It represents the possibility of an objective unity of experience, and is the complement of the subjective unity of experience, the transcendental unity of apperception. It is not thus a specific object ('table', chair, etc), but rather that which underlies their possibility.

Except for the word 'category', which we reserve for the former and not the latter use, we shall employ Kantian meanings of terms when within an explicitly Kantian context,
and the common, non-specialised meanings of terms when within a general context.
Part II

1 Cognition and its Relation to the Objective World

The notion that the form of our conscious perception of the external world is influenced by, or further, is delineated within the terms of the actions that we conjecture are successfully performable within it, is common to both a number of Twentieth-Century philosophical schools and to certain schools of cognitive science. The last of these we shall explore in the survey division dealing with the cognitive science context for cognitive bootstrapping. The current concern shall then be to directly relate this principle to the notion of cognitive bootstrapping in philosophical terms.

Of the philosophical schools that are relevant to our enquiry, it is most particularly the areas of phenomenology and ontology (respectively, the study of the relationship between mental acts and the external world, and the study of the basic categories of existing entity), along with their meeting point in hermeneutics (the study of the interpretation of meaning), upon which we shall choose to focus. This latter area of convergence is significant in that it provides a methodological apparatus for discussion of the interpretation of symbolic mental representations of the underlying physical reality within the context of the projected actions of the perceptive agent. We are consequently able to forge a connection between perception-action theories of cognitive agency, and the nature and grounding of symbolic representation.

Hence, we will seek to provide a philosophic grounding for discussion of cognitive bootstrapping by providing a descriptive basis of the internal and external (or equivalently subjective and objective, or mental and physical) accounts of a cognition in terms of the interaction between the distinct ontological realms. In this way, it becomes possible to envisage cognitive bootstrapping as providing a convergent mechanism for meaningful cognitive agency within the world, without having to invoke an a priori correlation between mental representations and the physical world (a position known historically as the 'Correspondence Theory of Truth', first formulated by Aristotle in the Metaphysics).

Following this, the discussion will proceed to attempts by recent thinkers such as Winograd and Mallory to relate the above considerations specifically to the field of artificial cognition, both in terms of the theoretical possibility of artificial cognition, as well as in practical proposals for the implementation of artificially cognitive systems.

Prior to this elaboration it is necessary to give an account of the classical view of cognition as it existed for twentieth-century philosophers following Kant’s 18th century revolution in metaphysics. In particular, we shall focus on the relationship between a priori knowledge and cognition in the Kantian system, and what implication this carries for the possibility of a self-founding perceptual system.

1.1 The Kantian View of Cognition

Kant’s philosophical struggle [64] was to unite the Enlightenment history of rationalism with the emerging empirical science of Newton as outlined in his Principia. David Hume [58], representing the Enlightenment tradition, had previously argued that knowledge is essentially empirical, that is, derived from a cognitive entity’s experience. This stance,
however, substantially contradicted several central tenets of Newtonian physics which postulated the existence of absolute notions such as infinitesimally divisible particulate matter, causality, fixed temporal and spatial dimensionalities and so on, all of which were deemed capable of persisting independently of cognition.

Hume had therefore recognised that there must therefore be a distinction between knowledge of particular cognitive patterns and knowledge of the concepts underlying those patterns, which (following Kant) are dubbed 'a priori'. For Hume this knowledge was derived purely by logical reasoning from absolute principles; they are, in Kantian terms, analytic and thus, finally, tautological, in the sense of being true by virtue of the definitions of the terms involved. Hume’s position (though not elaborated in precisely these terms), was that all a priori knowledge is analytic.

Kant sought to overturn this notion in his Critique of Pure Reason [64] by that arguing the critical concepts underlying our cognition, such as the law of cause and effect, were not analytic but rather synthetic. That is, the laws underlying perception are neither logically true nor logically false; they are, in a sense, constructed and must simply be assumed to be true in order for cognition to take place at all. (This notion was to later resonate with post-Galoisian mathematicians, who noted that the axioms in terms of which mathematical reasoning must proceed are implicitly chosen by mathematicians, and not derived by any prior logical process).

Kant’s quest thus became one of assessing how a cognitive agent may have synthetic knowledge that does not find its basis in empirical observation, in other words how cognition comes to incorporate synthetic a priori truths. His answer was essentially that of post-Galois mathematics, but applied to material reality; we cannot ever know the world as it exists in itself, independently of observation, because the faculties underlying cognition dictate the nature of our possible experience. We cannot, to employ one of his more famous examples, meaningfully conceive of an object that exists outside of our necessarily presupposed cognitive categories of time and space. Such an object may exist, but its being or non-being could never be empirically determined, only analytically conjectured. Since Kant regarded analysis as a form of tautology; a mere empty rearrangement of axioms, this was tantamount to denying metaphysics as an area of enquiry (metaphysics being the study of objects that exist beyond experience).

This has important implications for the extent to which cognition can freely experience the world, it that it denies the possibility of a pure, transparent empirical response to external objects (hence the title of Kant’s major work; ‘Critique of Pure Reason’, where the ‘Pure Reason’ in question was Descartes’s cogito that could deduce the absolute nature of the world from the basic fact of its existence; cogito ergo sum, ’I think therefore I am’). Human beings may well be capable, in Kant’s view, of learning many aspects of the ontology of the world, but the basic a priori categories such as the rule of cause-and-effect, or the topological connective character of temporal and spatial events could not ever be established by empirical experience, being rather their precondition.

There is, however, one important sense in which the Kantian cognitive entity resembles Descartes’s: the logical necessity of self-awareness as being prior to, and implicit in, awareness of any kind. Since self-modelling is implicit, at some level, in the auto-updating capabilities of cognitive-bootstrapping, and has also proved demonstrably use-

---

3Twenty-first century followers of this position were termed logical positivists, and enjoyed a short-lived flowering, representing the main stream of opposition to Kantian thought in that century.
ful in certain areas of autonomous artificial intelligence (particularly agent-based A.I., for instance [30]), it is useful here to discuss the nature of self-awareness and its relation to self-modelling in Kantian terms:

1.1.1 Transcendental Unity of Apperception

We have noted that non-analytic concepts not constructible from an agent’s *a priori* cognitive assumptions are categorised by Kant as metaphysical, and therefore beyond any meaningful discussion. However there is one important exception this rule; the phenomenon of *self*-awareness.

Hume had previously argued that the self is nothing more that the sum-total of its perceptions (the bundle theory of the self), in other words, that the cognitive self could never be experienced as a *single* perception, since the cognitive-self must necessarily encompass all of the possibilities of perception.

Kant’s concept of the noumenal world of things-in-themself not accessible to cognition, but nonetheless extant, gave him the scope to attribute a more substantial nature to the Humean totality of all (possibilities of) perceptions, with a greater emphasis on its character as a *unity*. Kant named this unity the ‘Transcendental Unity of Apperception’, the term transcendental referring to its, non-perceptible, noumenal nature. This unity of apperception he deemed to be prior to all other considerations in the construction of a cognitive entity, a requirement without which no *form* of cognition could exist. It is thus even prior to the synthetic *a priori* constructions that dictate the *nature* of possible perception. It should be noted that Kant does not rule out the possibility of a phenomenal, empirically discoverable ‘self’, such as a conception of ourselves as occupying a certain volume and being at a certain position, he rather asserts that there is a requirement in all acts of cognition that the self-sameness of the cogito (the ‘I am’) be discoverable across all possible unifications of the experience.

A strong reading of this principle would appear to deny any possibility of artificial cognition derived by means other than the evolutionary (a self-aware agent could not in principle be *constructed* by another self-aware agent because of the necessarily transcendent nature of the conditions for self-awareness). However, it does not rule out the building of a self-*modelling* machine, which would, to an observer, be behaviourally *consistent* with the notion self-awareness. For Kant, the issue of its *actual* self-awareness could not, even in principle, be resolved empirically, and has rather to be regarded as a metaphysical assertion.

At a weaker and more applied level, this principle would imply that a robotic agent with actions governed by, for instance, a finite-state transition model, and capable of learned responses could not, even in principle, *discover* that certain of its percepts corresponded to its own internal state-transition model, unless the capability to distinguish internal from external states was ‘hard-wired’ at the outset. Even then, the self-model so generated could not be fully isomorphic, given Russell’s limits on set self-membership. The issue is thus somewhat analogous to the halting problem for universal Turing machines, in which it is impossible for a machine that is emulating itself to be able to discover this fact, since this fact exists at a meta-logical level with regard to the logic utilised by the universal Turing machine.

The *a priori* and transcendental nature of the unity of apperception thus implies that
an artificial entity without a perceptual unity specified outside the immediate perceptual domain cannot even appear aware, let alone self-aware, with respect to a transcendentally unified observer occupying the same perceptual domain. Instead it would appear as a connected causal sequence of reactive events in space and time without any causal unity other than for the observer. Given, however, that such a specification does exist, we can now look more closely at the implied relationship between the cognitive agent and its cognitive object.

1.1.2 Objecthood

It is not enough for Kant that pure sensory impressions are united in cognition by virtue of its prior unity, rather it must additionally be the case that there exists an objective correlate of the subjective unity of apperception. That is, there must be synthetic concepts capable of describing the objective world that exists outside of cognition, and to which cognition refers; that is there must exist objects.

Kant argues that we cannot discover the possibility of objecthood (which is to say, the trans-temporal unity underlying the changing perceptions of the object) by observation alone - this can only be conceived a priori. Hence, cognition must include concept of objects (or at the very least the concept of the temporal persistence of identity) at the outset. We do not therefore perceive disconnected object impressions, but rather unified objects from a particular perspective. The object is then 'transcendental' since it can never be perceived in its entirety, being rather the unity behind the totality of perspectives.

It is consequently notable, and of central importance for the construction of artificial cognitive agents, that the object concept, as a unifier of disparate perceptions without being simply their summation (which would not be an explicit unity), serves to act as a compression mechanism for percepts. The a priori requirement of compact generalisation of sensory data in a conceptual (and hence non-singularly perceivable) object hypothesis is thus also a minimal a priori requirement of for artificial cognition; indeed any artificial agent employing a many-to-one sensory mapping, or percept classification scheme, fulfills this requirement. However, Kantian theory suggests that there must also be a transcendental subjective unity, which further requires that the objective unity tells us something specific about the subject. This is only possible if the agent deems itself to have a particular relation to the object concept by virtue of its actual and immediate subjective experience. That is, the agent's perception must be relative to the proposed object unity; in this way the totality of agent perspectives has an a priori unity corresponding exactly with the unity of the object-concept: agent actions then serve to link the individual perspectives together.

We illustrate this point schematically in figure 1 by way of contrast with the Humean model of cognition; here the Kantian position of an a priori synthetic transcendental object implicitly serves to embody the cognitive subject as a perspective at a particular location, allowing the cognitive subject to perceive itself as an a priori unity (the space of possible perspectives), in a way that is not possible in the Humean conception (it is instructive to compare this figure with that of figure 6 in Granlund's cognitive architecture overview [47] for comparison with an artificial cognition-motivated viewpoint).

The principle of the transcendental unity of apperception therefore regards the continu-
Individual realisations of cognitive possibles

LOCATE cognitive subject as a particular perspective on the transcendental object.

=> Kantian subject is aware of itself implicitly in its representation of the object.

Humean Cognition

Individual cognitive possibles

Individual cognitive realisations

Individual realisations of cognitive possibles suggest nothing about cognitive subject experiencing them.

Cognitive possibles are unified in the transcendental object.

The object is transcendental because it cannot be perceived as a single percept.

Kantian Cognition

Individual cognitive possibles

Individual cognitive realisations

Cognitive possibles are unified in the transcendental object.

LOCATE cognitive subject as a particular perspective on the transcendental object.

=> Kantian subject is aware of itself implicitly in its representation of the object.

Figure 1: A Representation of the Distinction Between Humean and Kantian Cognition in Terms of Spatial Attributes
ity of perspectives on a transcendental object as itself a transcendental object, and one, moreover, that is necessarily prior to all cognition. Cognition is therefore inherently experienced as a localised, embodied perspective, that can neither perceive itself nor its totality of possibilities other than as an abstract concept. The unity of perspectival perceptions is thus always a transcendental a priori, but one that must nonetheless exist and give rise to the possibility of objects and perspectives. Kant hence sets out a framework in which it is possible for a cognitive agent to view the world only in terms of its set of prior perceptual assumptions but which nonetheless realises the possibility of objectively validating higher-level cognitive categories built on these perceptual assumptions in a manner not available to either the classical empiricists or to the Cartesians.

Kant’s world, in summary, is thus defined by two realms, the phenomenal world of experience and the noumenal world of things-in-themselves, inaccessible to experience, but whose existence (if not nature) is guaranteed a priori by the fact of cognition. Thus, in contrast to the Cartesian model in which they are absolutely distinct, the Kantian conception of cognition allows for a subjectivity and objectivity that are co-depndent, and in which the cognitive subject is necessarily differentiated from, but embedded within, the objective world. This must then be the starting point of our understanding of the embodiment of cognition, and the minimal conceptual requirement in terms of which we approach the subject of artificial cognition and cognitive bootstrapping.

1.1.3 Kant and Cognitive Bootstrapping

It might first appear that the strong Kantian emphasis on prior synthetic knowledge leaves little room for the ontological learning required of cognitive bootstrapping. However, this is not the case; certainly an a priori ontology of object-cognition relations must be assumed by Kant before more detailed percept categories can be experienced. These percept-categories, though, if sufficiently law-like (such as, for instance, the tendency of spatial percepts to retain their internal morphology under translation) can then, in turn be used to derive a more complex cognitive ontology on an empirical basis - a cognitive ontology, in this case, of object relations. Thus a reasoning entity capable of embodying novel analytic and and synthetic categories is, in theory, constructible, provided that a minimum underlying a priori ontology has been assumed. The very possibility of empirical validation of percept/action categories thus rests on an a priori foundation of synthetic categories; it is these that embody the agent action potentialities involved in postulating a percept-action hypothesis validation experiment of the form:

If perceptual hypothesis $H_1$ is true then performing action $A_1$ will result in observation $O_1$.

If perceptual hypothesis $H_2$ is true then performing action $A_2$ will result in observation $O_2$.

(where the $O_n$ are the observational states, or perhaps stochastic distributions over observational states.)

Thus, while it is, in principle, possible to doubt the current percept hypothesis $H_1$, it is not possible in principle to doubt the causality implicit in the meta-notion 'performing a particular action results in a particular perception'; this is a priori and must be assumed before the possibility of empirical validation comes into being. Thus, in general,
while an autonomous cognitive agent may be free to reinterpret the world in the sense of being able to make an arbitrary choice of hypothesis, $H_n$, by which the world is to be interpreted, it is not free to choose an alternative set of action primitives to \{A'_n\}, or an alternative set of sensory primitives to \{O'_n\}, upon which the higher-level \{A_n\} and \{O_n\} are based, and in terms of which the perceptual validation criterion is constructed (and without which the perception-action mapping would be of entirely arbitrary construction)\(^4\).

For Kant, cognition is thus logically prior to cognitive bootstrapping. However, philosophy following Kant in the Constructive Empiricist [139] school has tended to modify the above position with respect to the percept-action relationship, noting that the noumenal world is capable of being defined existentially and phenomenally in terms of the occasional failures of the percept-action mapping. In other words, the world as it is in-itself expresses its nature in terms of the ‘resistance’ it supplies to our actions (and their expected perceptual correlations). The percept-action correlation is thus a translation between differing ontological frameworks, in the form of ongoing dialogue of identity between the cognitive and the environmental that is capable of virtually endless refinement. Thus, for the constructive empiricists, cognitive bootstrapping is logically co-equal with cognition.

Further evidence for this position from the cognitive science field comes from Winograd and Flores [146] who argue, in common with Mead [84], that conscious attention comes about when the already fully automated perception-action cycles fail to achieve their aim. Self-cognition, in the form of a self model is hence also required in order to be able to assess whether one’s perception-motivated actions have resulted in the expected perceptions: a cognitive agent must be able to perceive its actions and perceptions from a ‘third-person’ perspective. As Modayil and Kuipers [92] put it; 'bootstrap learning [is required] to move from egocentric ... sub-symbolic descriptions to symbolic object-based description’.

In other words, the trend has been to make the implicit possibility of cognitive bootstrapping present in the Kantian system an explicit necessity. Cognitive bootstrapping in this sense is thus the ‘feeding-back’ of an inferred action hypothesis into the percept space in order to refine the hypothesised percept-action relationship on the basis of the a priori motor and percept relationships. Post-Kantian metaphysics has thus come, increasingly, to resemble Bayesian statistics, a parallel that will be discussed from the mathematical point of view in a later section.

The first philosophy school in the Kantian mold, however, to fully explicate the relationship between phenomena and noumena as a cognitive bootstrap is hermeneutics, the study of meaning, which we are now in a position to address:

\(^4\)Obvious candidates for \{A'_n\} and \{O'_n\} in human cognition are, respectively, the motor complex and the space of visually-determined body-relative positions.
2 Hermeneutics: The Hermeneutic Circle and Cognitive bootstrapping

2.1 Textual Hermeneutics

A very full disquisition on the implication of hermeneutics for the goal of computer comprehension is given in [77]; we shall here concentrate on those aspects that are relevant to cognitive bootstrapping.

Hermeneutics emerged initially as the branch of philosophy that deals with textual interpretation, only later acquiring its interpretation as the branch of philosophy that concerns the mechanism of human understanding. As regards the former phase, modern hermeneutics can be considered to have begun with Wilhelm Dilthey's *methodological hermeneutics*, which sought to place texts within the context of their production in order to arrive at a scientific account of their meanings. The meaning of symbolic terms is thus dependent upon the *embodiment* of the symbol-manipulating agent within the objective world; it cannot be conferred by the manipulation of symbolic entities in their own right (without descending into semantic tautology, such as when attempting to derive the meaning of every word via dictionary definitions).

When this methodological approach is translated, as in later school of hermeneutics, from the textual domain to the more general domain of human understanding, the necessary situating of the symbolic understanding in the objective world will serve as a further refutation of Cartesianism, in particular of the notion that cognition can be defined independently of its actions in the world. Percept and action are thus codependent upon each other in a manner that goes beyond even the position of the constructivist empiricists. However, in the former, Diltheyan, terms of reference this principle of the embodiment of meaning states that, in order to interpret a body of texts (for instance the corpus of a single literary era), one must first understand all of its component parts. However, in order to correctly interpret each of the constituent texts one has first to be cognisant of the meaning of the whole corpus of texts.

This seeming paradox is reminiscent to those resulting from 'tangled hierarchies' in the domain of logic (of which many illustrations are given in [55]). Overcoming the paradox involved, for Dilthey, an initial *postulation* of individual meaning for each of the texts, from which composite meanings for a whole body of texts can then be built-up. These composite meanings can then be tested for mutual consistency, and any incompatibilities eradicated. In the light of this new, composite theory the original texts can be reinterpreted to form a new set of composite meanings, and so on.

This paradox-resolving movement, from part to whole and back again, Dilthey hence described as the 'hermeneutic circle'; the first explicit description of the mechanism underlying cognitive bootstrapping. Bootstrapping is thus required in order to arrive at a concrete theory of textual meaning given that we could not proceed, in the conventional logical manner, from initiatory textual meaning components, since the meaning of these components (the individual texts of the corpus) are themselves subject to the final interpretation arrived at with regard to the entire corpus; the perceptual 'zeitgeist' of a particular era (as expressed by the summation of its texts) is required order to interpret any individual text.
The 'hermeneutic circle' of interpretation is thus to simply propose any \textit{a priori}-plausible initial set of component meanings (for instance, a core set of words of known ancient-meaning with modern-day meanings attributed to the remainder), and carry out a reading of the entire corpus of work in order to arrive at an overall interpretation. This collective understanding is then utilised to \textit{reinterpret} the component texts in the context of the whole. These reinterpreted component texts are utilised to arrive at a new extrapolated interpretation of the corpus. It is hence tacitly understood (though not explained) by Dilthey that this endlessly reiterated process will achieve a degree of convergence on a final, stable meaning set.

Dilthey thus proposes the progressive, iterative testing and refinement of \textit{hypotheses} of interpretation in relation to a fixed set of texts (although explicit expression of the hermeneutic circle in these Popperian [108] terms was not to occur until Ricoeur). The total corpus is hence understood to have sufficient constraint information to fix its own meaning given an initially defined subset, in the same way that we can learn the meaning of every word in a dictionary on the proviso that we have independent lexical knowledge sufficient to understand at least a 'critical mass' of word definitions.

It is thus understood implicitly in the interpretative process that at least some word meanings are known relatively well in advance; our initial meaning hypothesis should be at least partially accurate, or else contain a sufficient quantity of \textit{a priori} correct word meanings as to prevent divergence of the process along arbitrary lines. Dilthey assumes that the hermeneutic circle of interpretation is effective because of our partially-shared interpretative framework (both socially and historically) with the original authors of the historical texts.

2.2 Non-Textual Hermeneutics

In terms of the wider context of hermeneutics, concerned with human understanding in general rather than merely the subset associated with language, the analogous problem is how we arrive at our initial conception of the objective world, and moreover, how we come to be sure of its accuracy.

Kant's conception of synthetic \textit{a priori} concepts provides us with the beginnings of an answer; we can be sure of the partial validity of our initial world model because it is already implicit in cognition; the fundamental ordering concepts of time, space, spatio-temporal continuity etc, are not testable hypothetical propositions, they are rather given in \textit{terms of which} we construct our hypotheses. We may thus make conjectures about the nature and existence of a \textit{particular} object, which may later prove to be false, but we cannot question the existence of the possibility of objecthood.

Our initial hypothesis in any application of the hermeneutic circle of interpretation to our understanding of the world in general will thus always be \textit{in terms} of the guaranteed validity of our immediate temporal/spatial perception components. Hypothetical conjectures are thus constructed \textit{in terms} of the extrapolation of these conceptions. (Hence, I can form conjectures about the particular scene I would view on opening a door, but not about the \textit{spatiality} of the area on the other side of the door, upon which this depends). Conjectures are thus always grounded by the invariants of the \textit{a priori} cognitive constraints: convergence of the hermeneutic circle of human cognition is thus always a possibility; the constraint information implicit in (for instance) a fixed
spatial dimensionality and topology is sufficient to prevent divergence in the range of possible hypothetical constructs. It is hence in this sense that the hermeneutic circle (as an activity rather than a tangled hierarchy of definition) is a mechanism of cognitive bootstrapping.

Within these later schools of hermeneutics that deal with the wider field of human understanding, two distinct branches, in particular, are evident: ontological hermeneutics and phenomenological hermeneutics (which relate to their two respective philosophical schools). Ontology, the study of things as they exist in themselves, we have discussed in Kantian terms above. Phenomenology, in contrast to ontology, purports to study the essence of things as they appear to cognition. First introduced by Edmund Husserl in his ‘Ideas: A General Introduction to Pure Phenomenology’ [60], phenomenology thus seeks to intentionally ‘bracket’ the problem of the nature of actual existents in order to arrive at the pure study of sensation).

In regard to hermeneutics these two schools of thought are exemplified by the respective figures of Heidegger and Ricoeur. They both seek to comprehend the relationship between mental representation and the world in terms of their interaction: however they differ in terms of the final grounding of the interaction.

For Ricoeur, in refutation of the Cartesian account, cognition is essentially embodied in the world as a consequence of its self-hood; cognition requires both self-identity across space and time as well as causal potency in order to exist at all. These conditions can only be fulfilled by an agent constructed of the same matter with which it interacts.

Ricoeur thus identifies a hermeneutic arc (cf [41]) that extends the notion of the hermeneutic circle introduced by Dilthey. This arc is then a two-fold movement that incorporates the hermeneutic circle of Dilthey as its first moment, corresponding to the transition from internal representation to objectivity. As the second moment of the arc, representing the transition from objectivity to subjectivity, he identifies the structuralist account of the (not necessarily just linguistic) world which that relies on the pre-understanding (ie a priori understanding) of the agent for its interpretation. Thus a purely positional account of the distribution of matter in a room only gains meaning when interpreted in terms of an agent with certain potentialities. For instance an object such as a screwdriver gains meaning by virtue of being a tool that may (via an embodied causal agent) act upon other object in a particular manner. An agent with a very different casual capacity (for instance one lacking hands) would necessary perceive the screwdriver in a different fashion. By taking the objective description of the world (as near as we can approximate it) as a starting point, we can arrive at a description of ourselves as an embodied percept-action complex (of the kind that is necessarily not apparent to us in purely percept terms). The hermeneutic arc is thus an alternating movement between percept-action hypothesis formation/testing, and self-modelling in terms of the ‘objective’ world in order to arrive at a description of our casual capacities for further hypothesis formation.

This is hence, once again, a true cognitive bootstrapping theory; however, unlike the hermeneutic circle idea of Dilthey, self-modelling is absolutely central to the process.
2.3 The Necessity of Embodiment

For Heidegger [53] this tendency to regard perceptions as being supervised upon by action possibilities reached its apotheosis. He proposed, in his ontological hermeneutics, that one’s sensations are completely defined by one’s acts and one’s possibilities. Heidegger thus envisaged our immediate sensation as being based on instrumentality (Vorhanden), in which for instance, our perception of a pen would be fully determined by our possibilities of using it, in particular our possibility of using it to write, with further social and contextual signification resting on what we may choose to write. Thus, the entirety of our being is employed in the perception of the pen, rendering the notion of an abstract mental plane of representations outside of perception entirely superfluous.

This notion also extended to the derivation of the objects of science (Zuhanden) from the praxial knowledge of action. Objective knowledge is thus abstraction from practical knowledge, and not its precursor. Relative to Ricoeur, Heidegger’s position is hence the stronger of the two, in that it argues that the totality of performable actions represents the fundamental mode of being of the intelligible world, rather than merely its appearance. In asserting that knowledge is intentional, there is hence a complete rejection of the notion that knowledge is representational; this is merely an artifact of dualistic Cartesian thought that falsely separates the body from the self in order to build a theoretically sound model of the world originating in the cogito ergo sum.

Ultimately, like his followers Sartre [119] and Merleau-Ponty [85], Heidegger was lead by this reasoning to deny the possibility of bracketing a perceptually independent ontological existence altogether. The agent’s being, as opposed to the world’s thus becomes the fundamentally irreducible entity from which everything else (both their existence and their essence) is derived (being’s corollary, non-being, or absence, being responsible for the necessarily non-present possibilities that give sensations their overall perceptual content). Only when action hypotheses fail to account for our percepts, do we, as it were, ’stand back’ from our perceptions and form a concept of objective existent independent from our selves; in the usual run of things objects are transparent to us - we only perceive our own potentialities unless these fail to be realised as expected.

There is hence in Heidegger’s work, an implicit ontological hermeneutic circle, in which the phenomenological perception of our possibilities can be transcended in order to arrive at a pseudo-objective account of the world. However, unlike in Dilthey’s account, this process can never be fully achieved in order to arrive at the absolute ontological existent (such a final scientific account); we must still define objectivity in terms of our failed action hypothesis (or percepts). Thus, in Heideggerian terms, we would define matter as that which resists our attempt to move it, giving it a degree of independence from our potentialities. In no way, however, is it possible to eliminate all aspects of the immediate presentation of the object to our action-potentialities from the object’s description (as might, for instance, be terminologically discerned in the apparently scientifically-objective description of object inertia as being resistance to action, which is to say the concept of inertia is inherently agent-relative). The bootstrap process of accounting for the world objectively must thus begin with our own immediate sensations, the end-point of our endeavour being hence to eliminate this immediacy and agent-centricity at some (inevitably infinitely receding) point.
The ontological hermeneutic circle may thus only be asymptotically convergent: we can arrive at progressively more objective description of the world, but we never arrive at a fully objective ontology (which would, for Heidegger, equate to a description of pure being). Transcendence of the subjective view-point (such as constitutes the endeavour of science) is thus possible, but never fully completable. In Nagel's deliberately paradoxical terms [95] the final, but unachievable, goal of any objective account of the world is to describe the 'view from nowhere', that is, an agent-centric description without agent-embodiment.

It is clear that these philosophical accounts of cognitive bootstrapping in terms of the hermeneutic circle have a bearing on the nature and possibility of open-ended artificial cognition. We would therefore, in the next but one section, like to determine precisely what form these limitations take. First, however, we would like to consider the possibility of ontology-free cognitive bootstrap models.

2.4 Ontology-Free Views of Embodied Cognition and the Possibility of Cognitive Bootstrapping

The cognitive models we have addressed thus far have been characterised by their ontological assumptions; they have all been required to make a presumption as to the underlying nature of reality in order to either to account for the existence of cognition, or to posit a specific cognitive mechanism. This underlying nature must then constrain the possibilities of what cognition can be of.

Other schools, however, have rejected this notion and sought to build cognitive models without significant a priori ontological structure, two significant exponent of this strategy being Quine and Wittgenstein.

2.4.1 Quine

Quine proposed a form of ontological relativism [109] in which it is impossible to fully constrain the underlying nature of reality by empirical or sensory data (since an infinity of equally valid interpretations or consistent theories are possible for a given sensory input). Quine therefore views cognition's positing of external object-entities as having no necessary correlation with reality (as Kantian theory insists upon), being rather convenient assumptions that have served to link sensory data together in the past, and which have no guarantee of utility in the future.

Part of being a cognitive agent is thus to acknowledge the endless potential falsifiability its conceptions, which are thus a matter of historical contingency serving to favour one set of categorical assumptions over the infinity of alternatives. For a cognitive agent to hold any conception of the world is thus merely for it to have had a particular history; there is no possible principle underlying the formation of general principles from specific examples (such as Occam's Razor). If any such method has had any success in the past, this is purely a product of the cognitive agent's particularly circumstances (for example, there is no a priori reason why the world should be, at base, simple rather than complex, and hence no a priori reason why Occam's Razor should always be effective).

Embodiment of the agent at a particular time and place is thus, for Quine, the only
possible basis for cognition (cognition being the holding of generalised conceptions of the world). A cognitive entity without a history has no means for distinguishing between possible ontological concepts.

Quine’s view is hence, to a degree, incompatible with the idea of cognitive bootstrapping, since, without an a priori method of constraining possible cognitive categories in the light of sensory data, no non-arbitrary method of achieving a final convergence of the perceptual model is possible. Arguments against Quine’s view are given at various points in the Cognitive Science and Mathematics survey divisions. We thus, for instance, argue with Millikan [87], [88] and Levin [72] that there is an a priori argument to be made for Occam’s Razor, and hence with Kant for the existence of prior (non-empirically justified) ontological structure. We thereby provide a grounding for the convergence mechanism that constitutes a critical aspect of cognitive bootstrapping, which, otherwise, would be groundless and the mechanism consequently divergent.

2.4.2 Wittgenstein

An ontologically neutral approach to philosophy that is compatible with cognitive bootstrapping is that outlined in Wittgenstein’s Philosophical Investigations [150]. Here Wittgenstein argues that the meaning of any word is exactly its use in the language. Conceptions of mind, consciousness, cognition, etc should not involve the postulation of entities that are beyond actual experience, but rather refer to the way in which we use the terms in daily life. Meaning is hence dependant on the environmental context; a particular disposition of neurons in the brain, or even symbols within an abstract mental space cannot, in themselves, account for it. It is hence not thus possible to have a private language, since word meaning is established only by prior agreement between differing linguistic agents; indeed the concept of meaning (signification) as separate from the word (as sign) refers only to this wider social context of word usage by mutual agreement. Our understanding of a language hence grows and modifies with our understanding of the way people act in relation to it; what Wittgenstein calls the ‘language game’. A prior, formal abstraction of the rules of language (syntax, grammar) etc is thus not possible.

The process of learning a language is hence a cognitive bootstrap, in that we arrive at meanings only by performing experimental actions within an environment. The only way we know that our conception of the word ‘bridge’ is the same as another agent’s is to establish whether they act exactly as we would in relation to sentences that use the term. Only by repeated hypothesis formation and testing (which is to say by communicating) can we become confident that this is the case. Any initial language conjecture can be used to bootstrap the process; we might subsequently discover that our use of the word ‘bridge’ corresponds to that of ‘crossing’ in the other agent’s usage, but, because of the similarity in meaning, this can only be established after considerable interaction in a variety of situations.

This bootstrapping process can reasonably be called cognitive in Wittgenstein’s ‘language game’ conjecture because the notion of a private sensation is just as inadmissible as that of a private language. All apparently mental sensations (such as thoughts, feelings etc) are, for Wittgenstein, as much a matter of contextual definition as are words. There is hence, as anti-Cartesians have termed it, no ‘Cartesian Theatre’ in which exter-
nal impressions are presented to consciousness. Cognition is hence only ever a linguistic bootstrapping process between agents embodied within a common world.

2.5 Implications for the Achievability of Artificial Cognition

2.5.1 Theoretical Implications: The Very Possibility Of Artificial Cognition

The account of cognition given by Heidegger and Sartre might thus appear to rule the possibility of artificial cognition, seeing cognition as something irreducible, related to (or even equivalent to) existence, and thus prior to any particular existents (such as neuronal structures). Both also see self-awareness as inseparable from awareness in general, having an a priori relationship with it. This latter relationship, however, should be susceptible to simulation via the more ontologically-neutral concept of self-reference. We might thus not be able, even in principle, to create an awareness along the lines proposed by the phenomenologists (or at least the existential phenomenologists), but we might, at least, be able to emulate its behaviour via formal self-reference. An extended discussion of the relationship between self-reference, defined in formal (Gödelian) terms, and self-consciousness, defined in the existential phenomenological terms of Heidegger and Sartre is given in [142]. Wittgenstein, we have seen, would not regard this distinction between consciousness and conscious behaviour as linguistically (and therefore metaphysically) significant.

More recently, one of the more persistent critics of the idea of artificial cognition is Dreyfus [28], who argues that the Representational Theory of Mind (in which the mind performs permutations of representations of the outside world) fails to take account of the contextuality, relevance and holism of perception. Discrete, atomic symbolic computation cannot account for the immediacy of the the cognitive situation (in a similar vein to Searle’s famous Chinese Room counterargument [123] to the proposal of strong A.I.). He suggests that only embodiment can provide a semantics of ordinary meaning, which left to symbolic computation alone would collapse into merely empty syntactic considerations. Moreover, this syntax, even if it existed, could never be available to cognition without involving problems of infinite regress. Thus, there can be no ‘algorithm’ underlying cognition which we could isolate and implement; only the situated, symbol-manipulating agent with an actual, sensible connection to the world can be truly cogent. The world, in effect, provides the ‘being’ behind the insubstantial formal categorisations of mind. Artificial cognition might thus exist, but not in any systematically per-formalisable way.

Suber [132] makes the argument that if mind can be expected to emerge from computation alone, then we should reasonably expect that semantics can emerge from syntax alone. However the Löwenheim-Skolem theorem of the branch of mathematics known as model theory demonstrates that even syntactic specifications with an infinite cardinality are incapable of uniquely determine a concrete, existing model. A very large degree of semantic ambiguity would therefore appear to be associated with any finitely formalisable set of syntactic rules, with the corresponding difficulty that this implies for the grounding of any putative ‘laws of cognition’ without a corresponding embodiment.

A similar view is given by Winograd in [147] who argues that the fallacy of cognitive objectivism (the view that cognition can be tangibly formalised) is caused by overly
formal logical structure of early attempts at simulated cognition (for instance his own \textit{SHRLDU} algorithm, which is capable of passing the Turing test for intelligent behaviour provided queries are restricted to the very limited but complete ontology of it’s internally represented world of Platonic solids and their transformations). Winograd argues that formal completeness of the logical system in which an agent is embodied is never available to that agent as a demonstrable fact (this would, in effect, constitute a \textit{Gödel} proposition [46]). Instead the \textit{embodied} agent can only allocate finite and partial resources to comprehending the world. This naturally leads him to abandon the notion of formally closed ontologies in any world description given by an agent; world descriptions have only to be (and indeed can only be) locally, and not globally, valid. Thus an artificially constructed cognitive agent is feasible in practise, but must necessarily be of an open-ended design (although he was later with Flores to reject this possibility [146], arguing, in common with Smythe [128], that any physically existing device for formal symbol manipulation cannot have any intrinsic meaning outside of that given to it by a subjective, situated agent; hence a computer program a performs a ‘task’ with ‘goals’ only if we so designate it. He does not, however, discuss the possibly that goal-directedness might originate mechanistically via Darwinist imperatives: feeding, mating etc - eg. [87], [88]).

Hermeneutic considerations would thus appear appear to rule out the possibility of \textit{formalisable} cognition, any apparent mental semantics attributed to artificial cognition actually relying upon inanalytic or meta-theoretical notions of meaning. However, both of the preceding objections to the notion of artificial cognition may be met by considering only \textit{robotic} systems equipped with cognitive bootstrapping mechanisms. The factual embodiment of the robot in the world meets Dreyfus’s objections on the semantic grounding issue, and the inclusion of cognitive bootstrapping serves to overcome Winograd’s objection on the grounds of logical closure. (While the initial stage of the cognitive bootstrap would of necessity be logically systematic, for instance a perceptual ontology built-up from Platonic solids, there is no implicit necessity that it would remain so during the iterative phase. Cognitive bootstrapping requires that we subject the initiatory hypothesis to empirical confirmation: relative to the level at which we specify the \textit{a priori} cognitive structure, which must be logically closed, higher-level hypothesis can, if necessary, relax the conditions of model consistency. For example, it might prove useful for an embodied robot to employ separate, only partially consistent environment maps for different aspects of a navigation problem: the topology of space in which the map-hypotheses are \textit{tested} must be a formally consistent one, however).

Given, then, that it is possible, in principle, for artificial systems to overcome the paradoxes of embodied cognition via the hermeneutic circle, it shall be instructive to survey extant mechanisms for achieving this \textit{explicitly}. (Other sections are, of course, also concerned with the utilisation of the hermeneutic circle as a mechanism for refining meaning in the generalised, perceptual, sense; we here limit our attention to those implementations that concern hermeneutics directly).

2.5.2 \textbf{Practical Implications: Real World Implementation of the Hermeneutic Circle}

A direct application of the hermeneutic arc theory within a computational environment may be found in [143]. This work concerns the construction of a software system, CIRAS,
for the logging and analysing of confidential accident reports with a view to extracting the significant risk factors. As such, it involves a transition from the qualitative to the quantitative realm, from the subjective individual accounts to an objective quantification of risk (at no stage is there any plausible access to a 'ground truth' factuality because of client confidentiality). CIRAS hence initiates a cognitive bootstrapping process by which an accident investigator is presented with an accident report and asked for an initial accident summary in terms of certain pre-existing casual classes (for instance 'ignoring a danger signal'), constituting an objective—subjective phase. These classes are defined so as to be mutually exclusive and exhaustive.

The hermeneutic arc then moves into its subjective—objective phase by instigating a 'structuralist' account of proceedings and defining casual structure on a word-by-word level. Thus, at the software's prompting, the report is broken down (encoded) into individual subject/object/action substructures ('close-reading') relating to the overall categorisations, and implicitly supplied with a subjective 'probability' assessment that individual elements may be so categorised. The documents are then reread within this format as a whole and the overall classification scheme reevaluated in this light. If it is indeed found necessary (through a cumulative impression of the subjective probabilities) to modify the classification structure, a new set of criteria is selected and the process is repeated. The user can then determine at what stage convergence is achieved. Proof that these convergences are objective rather than subjective is supplied by concordance data between fully independent CIRAS users; the agreement rate is found to be 71.73 percent in empirical trials.

CIRAS might thus best be described as a 'computer-aided' cognitive bootstrapping methodology, which still, nonetheless requires humans as the mediating agent within the hermeneutic circle. The possibility of a fully physical implementation of the hermeneutic circle is a central concern of Bitbol [8], who argues that the counterintuitive qualities of quantum mechanisms can be explained in terms of a particular variant on the hermeneutic circle, the 'epistemological circle'.

Here, situated agents instead of freely forming symbolic representation of external reality as in classical physics, enter rather into a composite relational system. Symbol representations exist only in so far as the agent can unvaryingly sustain them. However, the only invariant of quantum-mechanical cognitive agents is their own self-sustaining dynamical organisation. Thus, their cognitive domain is not a representation of what is external and pre-existing, but rather a fraction of the environment that is co-eval with them, within which their organisation can persist.

Self-organised agents (which can be compared with the non-quantum autopoietic agents of Maturana and Varela [80] below) do not, then, possess a faithful picture of the world, but rather only perceive notions relating to maintaining their internal self-consistency in relation to environmental disturbances. Each individual is thus, in Bitbol's usage, an eigenbehaviour, or an attractor (in the non-linear systems sense) for the dynamics of the autopoietic unit.

Hence, a classically embodied system implementing Newtonian physics and Boolean logic (such as those constitutive of conventional robotics) could not properly implement the epistemic circle of cognition (since its internal ordering is externally specified); only a fully quantum-mechanical could ever achieve this. Classical systems, Bitbol argues, cannot thus be considered truly cognitive because they implicitly assume an unbriddably
dualistic relationship between subject and object, whereas quantum theory is inherently monadic\textsuperscript{5}.

Thus, while such a cognitive entity may conceivably evolve, it cannot necessarily be constructed, given quantum-mechanical uncertainty relations. As we have seen, however, it is important to realise that this, and other theories similarly pessimistic with regard to the possibility of building cognitive entities, does not necessarily preclude the possibility of achieving a behavioural isomorphism between cognitive and non-cognitive entities. The goals of artificial cognition might thus be more modestly formulated in terms of concrete behavioural attainments, rather than the abstract one of achieving 'cognition'. This is the mentality underlying the Turing Test [138], which sought to move away from arguable and abstract criteria of cognitive achievement in preference to human-supplied behavioural assessments (we know when we observe intelligent behaviour, even if we cannot define it in the abstract). Given Wittgenstein's 'ordinary language' conception [149] of meaning, this may in fact constitute the only meaningful option for cognitive assessment.

Other researchers, however, do not see any necessity for invoking quantum mechanics in order to implement the hermeneutic circle; compare the previous quantum mechanical description of cognitive bootstrapping with that of Varela's (not explicitly cognitive) description of Autopoiesis in a Newtonian, biological environment:

"Autopoiesis attempts to define the uniqueness of the emergence that produces life in its fundamental cellular form. It's specific to the cellular level. There's a circular or network process that engenders a paradox: a self-organising network of biochemical reactions produces molecules, which do something specific and unique: they create a boundary, a membrane, which constrains the network that has produced the constituents of the membrane. This is a logical bootstrap, a loop: a network produces entities that create a boundary, which constrains the network that produced the boundary. This bootstrap is precisely what's unique about cells. A self-distinguishing entity exists when the bootstrap is completed. This entity has produced its own boundary. It doesn't require an external agent to notice it, or to say, "I'm here." It is, by itself, a self-distinction. It bootstraps itself out of a soup of chemistry and physics."

(in [14])

Cognition in autopoietic agents is hence self-derived, and reflects the distinction of agent and surroundings implicit in the autopoietic process.

Other classically realisable implementations of the hermeneutic circle include that of Winston [148], who proposes a textual system that learns by formal analogy (rather than strict deduction on the basis formal content). Furthermore, it learns rules on basis of its prior experience (constraint description in understood domains being utilised to constrain descriptive possibilities in unexplored domains). Novel hypothesis are hence suggested by the analogical process and submitted to the experimental environment for

\textsuperscript{5}Penrose [101] also gives arguments against Newtonian models of cognition in favour of quantum cognitive models on the grounds of our apparent, (though generally disputed), super-Turing computational abilities connected with our ability to comprehend Gödel's theorem. A counter argument employing similar meta-logical argumentation but asserting that Newtonian (or Newtonian-like) physics can can implement cognition is given by [99]. In this case, the requisite hyper-computation is made possible by the exception of the Newtonian class of physical models from standard set-theoretical constraints, such as in postulating the existence infinite fields of real numbers.
testing. Flexibility in the degree of matching insisted upon between the novel experimental domain and the previously experienced and understood domains are essential to the learning process; an analogy is only an analogy to the extent that it is not an isomorphism. Hypotheses hence serve, as much as anything else, to constrain the parameter freedom of the novel domain’s search space.

Hence, in so far as this process involves an initial (possibly arbitrary) assumption of formal categorisation, and furthermore has the capacity to reevaluate its initial categorical assumptions on the basis of new knowledge developed in other domains via analogy with this initial domain, the system may be regarded as implementing a hermeneutic circle (although Winston does not describe it in these terms). Critically, the system is capable of evolving toward a convergent world understanding irrespective of the prior assumptions with which it is initiated by virtue of having the possibility of questioning those assumptions in the light of later knowledge (even though this knowledge is built on those assumptions).

Like all true hermeneutic circles, it is the existence of ontological interdependence between part and whole (or in this case, form and content) that permits the possibility of bootstrapping an 'objective' set of perceptual states.

A somewhat similar approach is proposed by Bobrow and Winograd in [10] who, in addressing the concerns highlighted in the previous section, propose Knowledge Representation Language as a mechanism for capturing relations in human speech. Here, entities are described in terms of their relationships to other entities (prototypes), which can be understood as providing a perspective on the item under consideration. Typical properties of prototypes (which are effectively class medians) are assumed in the absence of observational information; however further observational information serves to reevaluate the prototypes in terms of which the observations are themselves made, thereby implementing a bootstrap cognition process. The system thus attempts to allocate finite computational resources effectively in order to determine an appropriate non-formal logic for a situated agent, which may thus be considered to constitute an embodied instantiation of the hermeneutic circle.

2.6 Summary of Division

We have thus, in this survey division, determined that the hermeneutic circle of perceptual meaning constitutes a particular instantiation of our notion of cognitive bootstrapping, in which the originary assumptions behind perception are subject to reinterpretation in terms of perceptual experience built upon those same assumptions, subject to the constraints imposed on cognition by the a priori Kantian laws that underlie the possibility of meaningful, and convergent, cognitive reassessment. We have also looked at the feasibility of implementing particular instances of the hermeneutic circle in the real world, detailing several specific approaches.

In the next part we take a more applied view, and look at cognitive bootstrapping from the perspective of cognitive-science.
Part III

3 General Cognitive Science

3.1 Introduction

Cognitive science may be defined as the division of science specifically concerned with mechanisms of cognition; as such, it touches upon cognitive psychology, cognitive linguistics, computer science, neuroscience and the philosophy of mind (in so far as it is empirically verifiable). The central notion that distinguishes this part of the survey from the others is hence this notion of *empiricism*, specifically the possibility of verification (or at least falsification) of specific mechanisms of mind.

In the following sections we consequently focus on the two broad divisions within the subject of cognitive science, the symbolic and the connectionist, detailing some of the manifestly cognitive-bootstrapping mechanisms that appear within the two subject divisions. We follow this with a section focusing on the necessity and significance of *embodiment* within cognitive science.

This will then provide the basis, in the final section of this division, for our predominant concern within the field cognitive science; *computational linguistics*, which encompasses both symbolic and connectionist concerns, providing models of perception that are both emergent and formally representational, but which maintain empirical objectivity by virtue of having arisen in the context of *communication* between agents.

The consequent focus in this latter section on the capability of cognitive bootstrapping mechanisms to generate linguistic *meaning* in a contextually refinable fashion is hence intended to complement the discussion of hermeneutics that took place in the previous philosophical division of the survey.

3.2 Approaches to Cognitive Science

A central concern within cognitive science is to determine whether human mentation is to be interpreted as the action of a large collection of individual computational elements (neuronal models, derived from physiological knowledge of the human, mammalian and reptilian brains), or whether it is to be interpreted at a higher level in terms of representations or schema. These two schools are respectively labelled the *connectionist* and the *symbolic*.

This distinction of approach is perhaps best reflected in their respective attitudes towards *simulation* of the human mind, both within the field of cognitive science as well as in the correlated engineering discipline of machine learning. Simulation of mental states is thus carried out either via emulation of large numbers of individual neurons, in which case we expect mental properties to arise as *emergent properties*, or else the simulation is executed at the schematic or representational level, in which case the actual underlying computational mechanics are of no inherent significance. In the former case, simulation is independent only of the underlying computational *substrate* (a logical unit can equally well be enacted by a radio-valve as a transistor), in the latter case simulation is independent of the particular computational implementation of the representational
algorithm.

The cognitive bootstrap mechanism identifyably exists within both of these sub-disciplines of cognitive science, the particulars of which are respectively outlined in the following two subsections.

### 3.2.1 Cognitive Bootstrapping in Symbolic Accounts of Human Cognition

A central problem for symbolic interpretations of cognitive psychology is to capture the fact the mental formalisms must be simultaneously both *computational* and *representational*; that is mental symbols must be manipulable by logical rules and also capable of referring to aspects of the world. Newell and Simon [98] were the first both to posit and to propose a solution to this problem from the perspective of cognitive psychology, centring on the concept of *physical symbol systems*. Here, physical relations (proximities, causalities and so on) provide the referential basis for symbol structures expressed within the brain.

Environmental adaptation (through Darwinian natural selection) is consequently the assumed agency constraining the formal symbol structure to mimic the physical environment (or at least those aspects of it that are relevant to the survival of the symbolic agent) within the Newell and Simon model. This aspect of the symbolic account was further brought out by Pinker and Bloom in the context of language evolution [106], who argued that 'grammar is a complex mechanism tailored to the transmission of [physically representable] propositional structures through a serial interface', the serial interface being the vocal communication channel.

It would therefore appear, in such physically-based accounts of symbolic causality, that the *representativity* of mental symbols is characterised by their capacity to ensure the continuing existence of the symbol-manipulating agent (or at least its genetically-contiguous progeny). Thus, while the symbolic manipulation system may be completely formal, the representativity of the symbols in the symbolic account is contingent and environmentally determined.

In this wider context, the particular symbolic model proposed by Newell and Simon can then be considered explicitly one of cognitive bootstrapping in the sense that world-model updates are achieved via genetic variations through mutation or sexual reproduction (equating to the hypothesis updating stage of cognitive bootstrapping), and are checked for their referencing ability by empirical practise in terms of the agent's attempts to survive within the environment (the hypothesis verification stage). The initial *a priori* symbol set is thus perhaps arrived at contingently, but the reinforcement learning of the symbol reference system will rapidly remove all traces of its random origin, until an appropriate representation is convergently found.

The above model assumes a relatively constant environment in relation to which the organism in question evolves. Conversely, where environments are not constant, and are changing at a faster rate than genetic adaptation can allow for, we would expect to find that the innate symbols acquire an inappropriate reference (such as, for instance, amongst humans, where animal threat assessments are calibrated to our hunter-gatherer past, rather than our urban/agrarian present; notably, the human instantiation of the primate's innate fear of the larger carnivores). It is therefore important, if Newell and Simon's notion of physical symbol systems is to be extended to symbolic inference mech-
organisms capable of autonomously updating themselves, that the Darwinian mechanism of bootstrapping be replaced by a more rapidly-updating technique that nonetheless retains the former mechanism's groundedness in the environmental survival imperatives of the cognitive agent: this shall be the subject of our section on the cognitive linguistic context of cognitive bootstrapping. We note for the present, however, that the innate, naturally-selected physical symbol set serves very effectively as an initial perceptual meaning hypotheses for cognitive bootstrapping.

3.2.2 Cognitive Bootstrapping in Connectionist Approaches to Cognitive Science

In contrast to the formal mechanics of the Symbolic approach, Connectionist accounts seek to comprehend meaning in terms of the aggregate information processing abilities of arrays of neuronal units, in intentional replication of mammalian or reptilian brain physiology. Cognitive properties can thus arise emergently, without explicit formal structure.

The most significant early demonstration of this neuroscientific approach was the perceptron model of neuronal activity given by Warren McCulloch and Walter Pitts in 1943, which set out a logical calculus implicit in nervous activity. This was soon modified by Donald Hebb [52] to include the possibility of strengthening the connection between the neurons as a product of activity, giving (in Frank Rosenblatt’s interpretation [116]) the activation rule:

\[ X.W + b > 0 \]  

(1)

\( X \) the input vector, \( W \) is a vector of weights, \( b \) the bias).

And also the weight-vector update rules:

\[ \forall n : \]

\[ W(n) = W(n) + [T - O].X(n) \]  

(3)

\[ b = b + [T - A] \]  

(4)

\( T \) being the anticipated output and \( O \) the actual neuronal output.

This capacity of the Hebbian perceptron to iteratively update itself might appear to permit it to be, in a limited sense, regarded as form of cognitive bootstrapping. Indeed, in so far as the perceptron may be regarded as cognisant of it inputs, the iterative updating of the weights and biases can be envisaged as the projection of a particular perceptual categorisation hypothesis (the current \( W \) and \( b \)) back into input domain (the objective space) such that an error can be computed from the disparity between the known class categorisations of the training data and those that occur under the proposed perceptual categorisation hypothesis. This error then determines the next perceptual category hypothesis and so on.

The distinctive movement between the percept and object domain thus resembles the more elaborate cognitive bootstrap mechanisms we have examined elsewhere, with the...
initial bootstrap hypothesis of relatively little significance, being subject iterative convergence on the final optimal model. However, the Hebbian perceptron is presented with unambiguous and unquestionable feature-categorisations (from which hypotheses are constructed), which it is not to free interpret (being, in effect, imposed as external constraints), and thus does not having the open-ended, potentially paradoxical capacity to assess the validity of its low-level percepts in terms of its learnt high-level percepts, as is the case for true of cognitive bootstrapping. This position would, of course, change were the system capable of forming additional percept hypotheses in a hierarchical fashion; however this is not the case for the unmodified linear perceptron.

Critical to the success of the perceptron model, this above-mentioned ability to converge upon an optimal, error-minimising W and b allows spontaneous generalisation of input data to occur; the first demonstration of human-like learning behaviour within the Connectionist regime, and a crucial benchmark of progress. However, Minsky and Papert were to point out [90] that this neuronal form, though capable of many classification tasks, was not yet capable of implementing the XOR Gate, and hence incapable of attaining computational equivalence to even basic human cognitive abilities.

The connectionist school of cognitive science hence remained subdued until the 1980s, when, via the Stone-Weierstrass theorem, it was demonstrated that multilayer perceptrons driven by backpropagation could in fact approximate any computational mapping function (provided the hidden layer were large enough). This discovery substantially renewed interest in neuronal-computational models of artificial intelligence, and along with this renewed interest arose the possibility of artificial cognitive models within the connectionist framework, and also the possibility of neural cognitive bootstrap mechanisms.

The most explicit such model is perhaps that of [79], in which the authors see Complementary Reinforcement Back-Propagation (CRBP) as way of directly achieving self-volitional behaviour in robots. Marshall et al. thus conjecture that self-directed learning behaviour comes about as the result of competing tensions, such as that between an agent simultaneously maximising its accuracy of prediction of future states while at the same time being compelled to seek out novel states. The 'homeostasis' thus achieved allows the network to bootstrap increasingly complex behaviour patterns. CRBP directly models this behaviour by, in addition to allowing backpropagation to reinforce internal goals in the conventional manner, also allowing the complement of the goal state, represented as a binary number, to serve as negative behaviour reinforcement during backpropagation. Thus the tension between contrary goal imperatives is directly modelled within the neural network structure, forcing the testing of cognitive models by deliberately seeking areas in which they break down, and hence refining them.

More broadly, in consequence of the hidden-layer network formulation, the now universal generalisation capability of neural network was to see it adopted as one of a number of default classifiers for use on arbitrary pattern recognition problems within the pattern-recognition community. The significant issue then became generalisation performance.

An important secondary issue for such pattern classifiers is then assessing the accuracy of such generalisations of the training data without having first obtaining an exhaustive data set, which is generally only available for idealised mathematical cases. In common with the other classification mechanisms adopted by the pattern-recognition community, there are, in fact, a number of such methods, one of the more commonly employed being
Efron’s bootstrap [31]. This involves the resampling of the training data to generate new classification generalisations such that a generalisation error can be computed, with this approach first being applied to neural networks in 1996 [136].

The bootstrap in the method of Efron is thus required to address the apparent paradox of the training data being employed as both a determinant of the classifier model generalisation, as well as the means for assessing its success at generalisation. We hence, once again, see bootstrap refinement mechanisms being employed to overcome difficulties of logical paradox associated with self-assessing generalisation mechanisms.

While this method has wide application in artificial cognition (eg [20]), note, however, that we reserve the term cognitive bootstrapping to refer only to bootstraps that are utilised to arrive a set of cognitive categories by validation methods employing those self-same validation categories (for instance using a conjectured environmental model in order to manoeuver round that environment to check the model’s validity). Thus Efron’s statistical bootstrap method, even if applied to cognitive data, lacks the feedback process that reinterprets the input data in terms of the updated perceptual categories.

**Other Connectionist Cognitive Bootstrap Models**  Even without explicit modelling of the neural substrate, considerable insight can be gained into the information processing techniques employed by the mammalian brain by utilising the techniques of experimental psychology. One such area of investigation that involves an active cognitive bootstrapping mechanism is the meeting point between visual and haptic perception (eg [61], [120], [121]).

When a mammalian agent interacts directly with the environment, it implicitly updates its visual model of the environment by haptic contact, using the a priori certainly of touch data to reduce the amount of ambiguity present in visual data (particularly the ambiguities of binocular scene reconstruction). Moreover, it appears that the human brain achieves this in a Bayes-optimal fashion.

The cognitive bootstrap in this model is thus the use of visual perception to motivate sensorimotor actions such as those involved in grasping for an object in order to test the validity of those same visual perceptions. As before, the bootstrapping of an initial, partially representative model and the iterative convergence between percepts and percept-motivated actions hence acts to overcome the logical paradox inherent in a self-validated perceptual system.

More generally the concept of the perception-action cycle (eg [122]) implicit in these visual-haptic models can by seen as the most tangible basis on which to implement a cognitive bootstrap mechanism, and one which will underpin many of the robotic implementations of cognitive bootstrapping listed in the Computer Science Perspectives portion of this survey. Perceptions are hence seen as environmental hypotheses while actions are hypothesis validation steps, or more specifically, vision is understood as a hypothetical linkage between possible instances of haptic contact (such as in 3D object reconstruction), and vision-motivated actions test the validity (or at least consistency) of these models.
3.2.3 Convergence of Symbolic and Connectionist Schools

The Boolean-logic completion of neuronal models having been demonstrated via the multi-layer perceptron model, the question of Turing-completeness then arises. The issue preventing this being achievable can be demonstrated to be the lack of memory associated with individual neurons (as opposed to the neuronal network as a whole, which does exhibit memory capability). It was hence determined by Franklin and Garzon [37] that the McCulloch-Pitts net augmented with expandable memory is Turing-complete and hence capable of arbitrary formal-language manipulation.

With this demonstration the Symbolic and Connectionist approaches had, for the first time, achieved a demonstrable equivalence: specific examples of the implementation of symbolic schema within a neural environment can, for instance, be found in [23]. Some authors were still to reject the possibility of significant crossover (most trenchantly [35]), citing the inability of connectionist systems to express symbolic compositionality (the concatenation of existing representations to construct new symbolic possibilities). Gärdenfors [40], however, repudiated this argument via the construction of a propositional language system based on the theory of functional dynamics applied to (purely abstract) information states. A neural network that undergoes learning generalisation of the Hebbian kind in response to new information is thus shown to perform an inductive inference of the kind recognised in formal logic. Hence the symbolic account is dynamic and emergent, but not simply an interpretation of the underlying neural connectionist model; it has actual referential capability.

On the assumption, then, that a compatibility between connectionist and symbolists accounts does exist, one of the more significant attempts to describe how their co-occurrence actually manifests itself is given in [78]. Marr, asserting that an information processing system cannot be understood without considering the use to which this information is put, delineates three cognitive 'levels': the computational, the algorithmic, and the implementational. Level 1 is hence teleological 'mind' that deals with agent intentions; level 2 is the symbolic computation layer of cognition that systematises thought, and level 3 is the objective, neuroanatomical distribution of material in the brain. Interaction between levels 1 and 2 thus determines how percepts are formed.

Marr hence implicitly assumes that a hermeneutic (which is to say, a theory of meaning) can only arise as a result of the embodiment of an algorithmic system within the physical world, though without explicitly treating the notion directly. Hence, without a particular imperative to reevaluate its goals in the light of material 'resistance' at the implementational layer, Marr's 3-level interpretation is not yet a cognitive bootstrap model; however it is certainly not incompatible with the notion. This conception of the embodiment of cognitive systems as a means of grounding symbolic manipulation does, however, receive a full treatment within cognitive science, and will be treated in detail in the following sections.

Hummel and Holyoak [59] also provide a unified symbolist/connectionist model within an incremental learning context by supplying an instantiation of analogical learning, LISA (Learning and Inference with Schemas and Analogies), motivated by Hofstadter's [56] assertion that all relational reasoning proceeds by analogy. In particular, Hummel and Holyoak argue that a reasoning system must encode information both as specific, local neural connectives as well as preserving identities of relational structure across the specific instances in order to allow for generalisation and symbolic processing: they
achieve this via *neural synchronicity* which allows the system to express variable binding. LISA is thus a *self-supervised* incremental learning system that automatically infers relationships from specific instances (and which hence constitutes a partial cognitive bootstrapping architecture in so far as it is capable of unlearning its initiatory weight settings via the inferred relationships).

Another possible route of unification for the symbolic and connectionist approaches, involving a common model for both artificial neural-network classification functions as well as formal symbolic constructs such as verbal grammar, is to view brain cognition as a form of *compression*. This approach, first suggested by Wolf [151], sees the essence of cognitive agency within the world as being the ability to represent the varied mass of sensory information in a portable, compact (and thus, generalised) form. Hence, grammatical rules may be regarded as a compressed version of the *language*, and *classification* may be seen as a compression of sense-data. The *object concept* itself can be derived by the redundancy or commonality between stereoscopic, or multi-angular images (compare this with the Kantian notion of the object concept as a unifier of perspectives).

In animal cognition, the mechanism motivating this compression is Darwinian natural-selection; biological agents employing better generalisers (which is to say, better compressors) use fewer neurons to find food by encoding successful hunting strategies in the most general manner possible. Since such agents inherently require less food to sustain their smaller neuronal budgets, there ensues a 'virtuous circle' in which they stand a greater chance of surviving and reproducing than their less efficiently-compressing relatives. Progressive generations thus increasingly enhance the likelihood of agents with ever more economised cognitive capacities (which is to say efficient sensory compression mechanisms). Moreover, when the environmental requirements are not static (as, for instance, in the context of hominid evolution), the selection pressure is towards ever more *generalised* cognitive capabilities (which is to say towards mechanisms of ever more efficient compression of *non-specific* data).

This is hence a fully cognitive bootstrapping mechanism *6* - the continuous need of the species to which the agent belongs to compress general, previously unexperienced sensory data amounts to a process of *hypothesis formation*, since the generalisability of the compression must be tested by feeding the hypothesis back into environment to establish its usefulness to the agent (in a process of hypothesis verification). The agent’s percept categories hence become self-foundational. This is also a *hermeneutic circle* since, in Marr’s model, the agent derives the *meaning* of its actions in terms of the effect they have on the agent’s own perceptions.

There is, however, a potential paradox here. We have, in outlining these biological, cognitive-science derived cognitive bootstrapping models, furnished a scientifically-derived explanation for the existence of cognitive bootstrapping. In the philosophy section, though, we identified a cognitive bootstrapping-based explanation for *science* itself (that is, one in which the scientific method is itself seen as a formalisation of the cognitive bootstrapping that constitutes the fundamental mode of existence in ontological hermeneutic agents). Resolving of the paradox ultimately depends on our assumptions

---

*6*Interestingly, Graham Cairns-Smith [19] argues that the origin of life itself constitutes a bootstrap process in which an initiatory system of biological information storage and self-replication forms the basis for successive systems that finally culminate in the current DNA-based system that has supplanted all previous systems. Thus there is no inherent statistical problem associated with the spontaneous appearance of complex DNA molecules.
about the final ground of objective truth - whether it is the set of a priori assumptions that are necessarily inherent in cognition but which are necessarily not the subject of empirical enquiry, or whether it is the empirically-verified objects of perception themselves.

The scientific method, being observation-based, is ostensibly in the latter category, although scientific laws themselves (for example, the conservation of energy) must always be conceptual (in the sense of not being accessible directly to cognition; they act rather as proposed orderings of perceptions, with the ordering concepts, such as causality being essentially pre-scientific). The former category, insofar as it relates only to the a priori laws of perception, is potentially antipathetic toward the concept of strong A.I. (the possibility of intentional construction of cognitive agents) since in Kantian terms is only possible to intentionally build devices in terms of a priori cognitive concepts, rather than inclusive of these a priori cognitive concepts (though this does not preclude cognitive machines evolving spontaneously). Cognitive science implicitly favours the latter option (that the final ground of cognition is observable, or at least in observable in its consequences), and hence considers the scientific method and its final objects to be logically prior to cognition: we shall return to this point in the conclusion. For the purposes of the remainder of this survey division, however, we shall consider only the latter possibility.

Having established, then, that the notion of cognitive bootstrapping is broadly consistent with both the major schools of cognitive science, the connectionist and the symbolic, we can now proceed to focus on the specific issue of agency within cognition; in particular, we shall, in the next section, focus on the centrally important notion of embodiment.

### 3.3 The Significance of the Embodied Mind Within Cognitive Science

The notion that the form of our conscious perception of the external world is dictated by, or further, defined within the terms of the actions that we may perform within it, is common to both phenomenology and to several long-standing schools of cognitive science. (Dewey had argued as early as 1896 [27] that perception, thought and action must be considered as part of the same stratum. Thus, rather than first perceiving a scene, then thinking about its content, and only then performing an act, he argued that perception is actively modified by manipulation of the environment, in distinction to the classical notion that actions are dictated by perceptions).

If we were thus to divide cognitive science into the theoretical and practical schools, an example of action-based perception in the former school is given in the study of affordance, a term first coined by James Gibson [43], and specified in [83] as having the following properties:

- 1. An affordance exists relative to the action capabilities of a particular agent.
- 2. The existence of an affordance is independent of the agent's ability to perceive it.
- 3. An affordance does not change as the needs and goals of the agent change.
Affordances, being the action possibilities of the agent’s environment, hence overcome the dualistic subjective/objective divide in favour of a monadic account of perception. Affordances are objective to the extent that they are invariant to arbitrary shifts in interpretation, however, they are subjective in so far as they require an agent to provide a relative frame of reference.

Other theoretical schematisations of embodied cognition along these lines include Lakoff’s, cf eg [70], which argues that reason far from being an abstract logical discipline independent of the body, is in fact patterned by the way in which the spatial awareness of our body’s agency is constructed. Glenberg similarly argues in [44] that conceptualisation is constrained by the structure of the environment, our bodies, and our memory capacity. Hence, a key human skill is in forming a conception of the environment that is independent of the environment. Memory (of previous conceptions) is the key to achieving this (and may be related to our later arguments concerning the importance of conceptualising our own agency in effective cognitive updating).

In having sought to find a middle ground within the subject/object dualism of classical Psychology, Lakoff’s scheme consequently lacks a foundational ontology (ontologies being classically objective), such as those sought by fundamental science.

On the applied side of cognitive science are the searches for neural correlates of embodied cognition, for instance Berlucchi and Aglioti’s [6] argument that the imitation of movements within neonates is indicative of an implicit neural body-structure model from which later neural body-structure models are determined. Moreover, crucially, this model provides a reference frame that extends to the neural determination of inanimate object models. The mechanism of object understanding is thus a cognitive bootstrap to the extent that it requires, firstly, an initial set of a priori assumptions (the implicit model) in terms of which the world model is defined and, secondly, a constructive dialectic between the world and agent’s world-model hypothesis in order to refine this model.

This work, and others like it, thus serve to validate Piaget’s [103], [104] notion that higher cognitive functions have their roots in lower-level biological mechanisms.

A similar idea is expressed by Millikan [87], [88] with regard to language and intentionality. Here she argues that function can only be attributed to an entity within a biological context; the purpose of a leopard’s spots are to provide it with camouflage for hunting. Purpose is thus not defined by a particular agent’s mental state, or even it’s immediate environmental context, but rather its individual and species history.

Millikan hence proposes a biological solution to the Kripke-Wittgenstein paradox, which relates to the apparent impossibility (at least in Kripke’s reading of Wittgenstein) of establishing absolute conceptual or perceptual identity between agents, since an unbounded notion such as ‘addition’ could never be proven to be the same for both agents. In this example, one agent’s rule of addition might hence be the ‘correct’ one; $\forall x, y z := x + y$, whereas the other agent’s rule might be some near approximation such as; $\forall x, y x < 5 \times 10^9, y < 5 \times 10^9 \ z := x + y$; else $z := 5$. In any reasonably finite scenario these agents would falsely form the impression that they both had the same understanding of addition. The paradox is that this impossibility would appear to reduce individual observations to the status of ungeneralisable atomic facts, in which case all concepts of mentality are illusory. Millikan’s resolution of the paradox is to propose that natural selection serves to remove the latter formulation of the addition rule on

42
the grounds of its inefficiency; it does the same essential referring as the former rule with regard to reasonably small numbers such as those the agents experience in their biological lifetime, but uses more computation to do so. Hence aggregate natural selection will favour the smallest generalisation consistent with the biologically necessary referents (thus providing a basis for Occam's Razor). Meaningful linguistic exchange between agents of the same species is thus possible.

Millikan's work thus overcomes the classical problem of reference, where the relation between percept and object appears to be arbitrary (why, for instance, do we regard a rabbit as principally a single entity rather than as a collection of organic sub-objects or as a subpart of a species-collective). She argues instead that the particular form the percept takes in relation to the object and the agent-object interaction has an inherent survival value for the agent (we have traditionally hunted rabbits for food, and so regard an individually huntable unit as a single perceptual entity). Percept models that do not efficiently model the survival-relative aspects of the object in relation to the agent's action possibilities simply cease to exist on an evolutionary time-scale.

Anderson [1] also argues for the Darwinian, non-Cartesian nature of embodied cognition, one that it is not fundamentally representative, but rather interactive, since 'the world is its own model' and therefore needs no perceptual states distinct from action possibilities.

Drawing together the various differing strands within cognitive science, Gallagher [38] provides an extensive interdisciplinary study that seeks to strike a middle ground between the physically reductionist accounts of brain cognition and Cartesian top-down approaches, acknowledging the cognitive science and phenomenology are inseparable sides of the same coin.

A similarly complete treatment of the concept of embodiment within cognitive science is aimed at in [42], which argues that 'Cognition is what occurs when the body engages the physical and cultural world and must be studied in terms of the dynamical interactions between people and the environment'. The embodiment of intelligent behaviour thus acts as a constraint that gives rise to the existence of human language and human thought.

With this in mind, we now turn to the specifically linguistic aspects of cognitive science, with a view to establishing the centrality of cognitive bootstrapping to this area and, in particular, its relevance to syntactic forms of symbolic representation.

4 Cognitive Linguistic Context for Cognitive Bootstrapping

4.1 Section Introduction: The Origins of Symbolic Representation in Cognitive Agents

We shall, in this section, focus on the evidence for the assertion that symbolic representation arises as the result of communication between cognitive agents (or even from self-communication in the case of a single agent with the ability to model itself). In particular, we shall seek to demonstrate that the cognitive bootstrap model, in serving as the mechanism underpinning the fundamental subject/object division (as well as
overcoming the philosophical paradoxes that have traditionally been associated with it), also provides, in the process, a symbolic and communicable representation of the world in terms of an intrinsic, but evolvable, language system.

4.2 Origins of Language in the Symbolic Interchange of Percepts

The basis for this argument is then that, in attempting communication with another cognitive entity, we must necessarily find a representation of the simultaneous commonalities of our experience. That is we must, in some way, abstract from our personal (percept/action-based) experiences in order to find that aspect of them that is accessible to a real or putative second entity also embodying a perception/action relation.

As we have seen, the possibility of the abstraction of aspects of our perception/action experience into the third person is, for Kant, already implicit in our perception of the world. Perceptions are inherently experienced as having a certain unifying constancy under the transformations associated with agent actions; that is, we perceive objects from perspectives, rather than pure sensory qualia\(^7\). The abstraction of our experience required for communication is thus implicit at the outset.

However, this rigid, predetermined ontological structure might not initially appear to allow for the possibility of learning a language, or for the spontaneous evolution of an appropriate language between cognitive entities attempting to describing their cognitive world at a greater level of detail (such as, for instance, constitutes the goal of the experimental science community). How is it then possible, in a communicative context, for cognitive entities to establish a common description of the world that goes beyond what is necessitated a priori?

Cognitive bootstrapping supplies an appropriate solution framework that does not falsify any of Kant’s constraints on the nature of the a priori cognitive categories. We have seen that new percept hypotheses can be formed by the cognitive agent in terms the existing Kantian categories, and ‘projected’ back into the perceptual environment as perception/action conjectures for empirical (that is, action-based) verification (or rather, consistency checking, since we cannot intrinsically confirm or deny the accuracy of the representational primitives). Successfully empirical testing of these novel perception/action conjunctions can then expand the range of the agent’s cognitive categories, which in turn provides a range of new perception-action hypotheses with which to recommence the hermeneutic cycle\(^8\) of perceptual reinterpretation.

What is important to appreciate in the current context, however, is that this mechanism constitutes a communicative (or self-communicative) process, in which communication serves as the methodology by which symbol and symbol-meaning are differentiated and defined. Implicit in the cognitive bootstrapping process is thus a division between subject (the symbol hypothesising agent) and environment (the entity capable of falsifying the symbol meaning hypotheses), both of which must be represented within the cognitive agent. The agent must therefore necessarily employ a distinct model of its own

---

\(^7\)The immediate and irreducibly-in-analytic aspects of perception, such as colour.

\(^8\)The hermeneutics in question being the link between the ‘semantics’ of the asymptotically-obtainable, observer-independent object ontology and the ‘syntax’ of the currently assumed percept-action relation.
perception-action relationship (framed in terms of the current best percept-action hypothesis), along with any new conjectured perception-action hypotheses, in such a way that they may be falsified against the current best percept-action hypothesis.

It is then in this sense that the method is a bootstrap; at no stage do we have access to the ground-truth object-centric percepts - validation must instead take place in terms of the previous partially-representative percept hypotheses. All that prevents the method from logical collapse into either tautology or relativism is the guaranteed validity of the Kantian a priori percept categories upon which the hypotheses are based (along with the assumption of Occam’s Razor as means of distinguishing between equivalent percept-action conjectures).

The final convergent percept-model of the cognitive bootstrap is therefore, if not the absolute Kantian noumenal world independent of the observer, rather the object (as opposed to the view)-centric description of the world that underlies the percept-action relation. That is, we obtain, either by communication with another cognitive agent, or, in the case of general cognition, by an implicit communication with models of one’s previous cognitive agency, a final, stable description of the world independent of one’s particular perspective. We hence achieve the seeming paradoxical position of transcending our particular perceptual frame in order to arrive at a symbolic description of the world that makes no reference to a particular point of view, and hence which approaches the level of objectivity assumed by the classical empiricists, who assumed that no form of perceptual mediation between cognition and the world can exist. All that is required for this to occur is that the initialisation of the cognitive bootstrap partially embodies this ideal: the a priori categories required in order for Kantian perception to take place at all are sufficient to allow this.

Justification for the above communicative account of how cognition arrives at an objective world representation is found within the subject of linguistics. Rohrer [115], for instance, suggests that linguistics should properly be regarded as a sub-science of cognitive science, proposing that the basis for language is the projection of one’s own agency model into the perceptual domain; that is, a relativising of experience in order to establish a common frame of reference, and thereby arrive at an object ontology. Perry [102], Bermúdez [7], Metzinger [86] and Baker [4], also agree that cognitive self-awareness (as manifested by a linguistic token equivalent to 'I) requires a communicative domain in which all communicating parties have internal object models of both the world and of the various inter-communicating agents; in no other circumstances can one explicitly attribute perceptions to oneself.

Pinker [105] further argues that language derives from an initial cognitive orientation (arguing, for example, that the fundamental noun/verb split mimics the percept/action division), which then develops along more complex lines via a semantic bootstrapping mechanism. The bootstrap proceeds by alternatively hypothesising and then (when sufficiently established) perceiving progressively more refined configurations of noun/verb and perception/action pairings.

Language development and cognitive bootstrapping would therefore appear to be inseparable. Critically, from the point of view of establishing such notions on a scientific footing, these cognitive-agent-based understandings of the process of symbol formation and language development lend themselves straightforwardly to empirical testing via computer simulation. There consequently exists a considerable body of literature that
deals with the generalised computation architecture of cognitive agents in which language and symbol formation is considered at least implicitly (cf eg Minsky et al. [89], Edelman & Intrator [29] for overviews). However, in the next section we would like to consider artificial embodied cognitive architectures that address the notion of language formation on its own terms:

4.3 Language Evolution in Embodied Artificial Cognitive Agents

4.3.1 General Ontology Learning Mechanisms

Before commencing with a description of spontaneous language formation experiments in embodied artificial cognitive systems, we shall briefly digress to examine the various mechanisms for generalised symbolic ontology learning that are applicable to simulated language inference.

Any system for learning, ab initio, the ontological relations of abstract symbol-forms underlying perception requires three key elements; a mechanism for generating hypotheses, a mechanism for hypothesis verification, and prior to both of these, a fixed logical structure. The most general such construction is a theorem prover.

However, given that we have, in the preceding section, established the principles of self-representation and embodiment of the cognitive agent within the world that it seeks to model as being at the origin of ontology, there then immediately arises the problem of computability. If, for instance, one were to take the most general case and utilise a universal Turing machine\(^9\) as the hypothesis generation mechanism in order to arrive at the particular Turing machine that describes the underlying mechanism behind every possible percept-action state available to the agent (generating a progressive series of hypothetical Turing machines of index \(i\) for empirical validation with respect to the set of 'input' percept states, \(P\)), we would eventually require (due to the necessity of self-modelling) that \(i = j\), where \(j\) is the index of the universal Turing machine utilised by the agent to generate the hypotheses. Assuming logical consistency, the system would then fail to reach a stable output solution in consequence of attempting to output its own percept-action (input-output) mapping (corresponding to the classical halting problem).

It is therefore necessary to limit the range of application of the theorem prover in some way, such as imposing a restraint on the range of the hypothesis generation (dealing, for instance, only with coarse-grained perception-action models, or, as discussed in the next survey division, limiting the temporal iteration budget). Otherwise, we must impose a restraint on the logical structure of the world model. This latter option is the most relevant to discussing language formation within cognitive agents, where the logic is typically limited to either propositional logic or else to first or second order logics. We summarise these briefly below:

\(^9\)A Turing machine capable of simulating any other. In particular, one that takes the input \(a\) to a Turing machine of index \(i\), and produces (for a given input \(a, i\)) the corresponding output \(b\). We write this as \(T_i(a) = b\) in the former case, and \(U_j(a, i) = b\) in the latter case (ie for a universal Turing machine of index \(j\)).

46
Propositional Logic  A propositional calculus is a subset of a larger logical system that determines how to form logical propositions from a set of axioms. It is thus a collection of inference rules for arriving at logical conclusions given a set of supplied facts. Consequently, it requires a formal grammar. Implicit in the grammar is a semantics, an implied reference to world ontology indicating the significant aspects of the deduction (for instance the differentiation between a logical connective and an atomic sentence is assumed to be more than merely syntactic). This ontology is, moreover, closed, in the sense that propositions not provable in the language are assumed to be false, a restriction that serves to ensure decidability.

In the axiomatic (as opposed to the syntactic) form of propositional logic, the single rule of logical inference in propositional logic is modus ponens (from the Latin, 'mode that affirms'), which permits deductive arguments of the form: If X, then Y. X therefore Y. (or in operator form; X, X \rightarrow Y \vdash Y).

Logical inference in an applied scenario under Boolean restrictions would therefore typically take the form of a syllogism of the type:

If the ground is cold or the air is cold or the sea is cold then the weather is bad.

The air is cold.

Therefore the weather is bad.

This inference system was the first to be formalised in a systematic fashion, being set out in Aristotle’s Prior Analytics within which is argued (incorrectly, as it later transpired) that every deductive argument can be expressed in this form.

What it is, in particular, that is lacking from this propositional logic form is the incorporation of variables. It is hence incapable of generalisation of any kind, lacking the capacity to adapt to novel situations (a key requirement for cognitive agency by almost all definitions of the term). To achieve this we must turn to first-order logics:

First-Order Logics  The most immediate way to develop a more complex logical calculus is to introduce additional axioms that are applicable to the finer distinctions of the sentential entities that occur in propositional logic. In particular, if the atomic sentences of propositional logic are divided into terms, variables, predicates, and quantifiers, they give rise to first-order logic, or first-order predicate calculus. First order logic is thus a theory within symbolic logic that permits the formulation of quantified statements of the form; "there is at least one X such that..." or "for any X, it is the case that...".

We could, using first-order predicate calculus, hence use the previous meteorological example to write the general rule:

For all entities X, if X is cold then the weather is bad.

A first-order logical system is thus capable of forming abstract concepts, rather than merely accreting aggregations of specific instances. We might thus expect an embodied first-order cognitive logical system to present a mechanism capable of adapting to entirely novel situations in a way that is impossible under a purely propositional logic calculus.

47
We shall see that this is true to an extent. Systems for inferring first-order logic typically deal with Horn clauses, which are implication conjectures in which there is a conjunction of multiple literals (elementary propositions and their negations) that lead by implication to a single literal. This is, in effect, a class allocation, and gives rise to the relative ubiquity of this approach within the field of logistical pattern recognition. Systems of this type are collectively described under the umbrella term 'inductive logic programming' (ILP) (see Muggleton [93] for an overview).

A typical problem structure might thus be the determination of biological species descriptors from specific examples; that is (to give one of Muggleton's examples), a determination of the general rules:

$$\text{class}(A, \text{reptile}) :\text{- has-covering}(A, \text{scales}), \text{ has-legs}(A, A).$$
$$\text{class}(A, \text{mammal}) :\text{- homeothermic}(A), \text{ has-milk}(A).$$
$$\text{class}(A, \text{fish}) :\text{- has-legs}(A, 0), \text{ has-eggs}(A).$$
$$\text{class}(A, \text{reptile}) :\text{- has-covering}(A, \text{scales}), \text{ habitat}(A, \text{land}).$$
$$\text{class}(A, \text{bird}) :\text{- has-covering}(A, \text{feathers}).$$

from specific background knowledge of the form:

$$\text{has-covering}(\text{dog}, \text{hair}).$$
$$\text{has-covering}(\text{crocodile}, \text{scales}), \text{ etc}$$

in conjunction with positive and negative examples of the type:

**Positives:**
$$\text{class}(\text{lizard}, \text{reptile}).$$
$$\text{class}(\text{trout}, \text{fish}).$$
$$\text{class}(\text{bat}, \text{mammal}).$$
$$\text{etc}.$$

**Negatives:**
$$\text{- class}(\text{trout}, \text{mammal}).$$
$$\text{- class}(\text{herring}, \text{mammal}).$$
$$\text{- class}(\text{platypus}, \text{reptile}).$$
$$\text{etc}.$$

Hypotheses are therefore equivalent to PROLOG programs, and ILP involves their inference from specific examples. Generally, ILP methods will adopt either a general-to-specific or a specific-to-general search mechanism for rule generation. GOLEM [93], employing the concept of 'least general generalisation' is an example of the former, while FOIL [111] corresponds to the latter category.

As PROLOG meta-programs, both of these techniques have in common that they effectively override PROLOG's closed ontology assumption (namely that inferences outside of provable range of the current hypotheses are necessarily false), and adopt instead a more open-ended approach to theorem building. As such, these methods find a number
of applications in real-world classification scenarios (for instance pharmaceutical side-effect prediction), but do not, in themselves, constitute the ideal solution to the problem of embodied cognitive ontology learning. In particular, they cannot explicitly infer new perceptual classes from pre-existing predicates, and so are not, of themselves, capable of cognitive bootstrapping.

As an example of this, suppose, for instance, that an ILP-based cognitive agent embodying a perception-action cycle is equipped with the *a priori* perceptual states, \( \{P_n\} \), and motor capabilities, \( \{M_n\} \). Suppose, further, that following a series of experiments involving randomised permutations of these action primitives, it has had cause to infer the novel PROLOG motor rule:

\[
M_a(o, x, y, z) : = P_1(o', x, y, z'), \ z' \text{ is } z - 1.
\]

(corresponding, intuitively, to the motor rule 'Object o can only be placed at position \((x, y, z)\) provided that this position is directly on top of some other object \(o'\)).

To an autonomously updating perceptual mechanism, inference of this motor rule ought also to suggest inference of the corresponding cognitive category:

\[
P_a(o, o') : = P_1(o, x, y, z), \ P_1(o', x, y, z), \ z' \text{ is } z - 1.
\]

\(P_a\) correlating with the concept 'is on top of', such that the originally inferred motor rule becomes \(M'_a(o, o')\) - intuitively, 'Put object o on top of object \(o'\).

In essence, we require that the cognitive system enact an elimination of redundant predicates in the inferred motor possibility in order to infer a new, higher-level percept-action correspondence that is always successful: i.e., generically, we require a P(X) such that M(X) is always valid (\(M'_a(o, o')\) does not require conditional satisfaction in the above). This is not possible via straightforward ILP. In order to accomplish this we need to consider at least the second-order logics:

**Second-Order Logics** The most commonly utilised logic in agent-based artificial intelligence is monadic second-order logic. Second-order logic differs from first order logic only in its ability to quantify over *properties* of variables. It is incapable of admitting a proof theory, and consequently always computable. More particularly, it is capable of implementing finite state automata (of the type capable of generalising over repetitive percept-action cycles), and non-deterministic pushdown automata (of the type capable of modelling Markovian systems with finite memory). In terms of Chomsky's language hierarchy [24], [25], second order logics can, respectively, define Type-3 regular languages and Type-2 context-free grammars.

Moreover, monadic second-order logics limited to Type-3 regular language expressions can be *optimised*. That is, finding the finite state machine with the least number of states capable of performing a given function is always *decidable*. Cognitive updating is therefore, for an agent employing this logical restriction on (finite sets of) observations, *always* convergent, and, furthermore, convergent in a finite number of steps. Common learning systems not explicitly conceived in terms of second-order logic can also have equivalent capabilities; Forcada *et al.* [36], for instance, show that discrete-time neural networks are capable of inferring deterministic finite automata (though, in practise, they would recommend discrete algorithmic methods as a solution mechanism).
It would thus appear that connectionist approaches can equally well implement constrained logical forms within cognitive updating, but yet capture patterns of perceptual inference rich enough to allow for inference of abstract perceptual forms. In general, when searching for a suitable logic for cognitive bootstrapping within real-world environments we have two principle requirements. The logic should not be so powerful that issues of non-computability arise (unless steps are taken to discard hypothesis generation at the level of complexity at which non-halting hypotheses are taken to occur [i.e after a certain number of iterations]). The logic must also not be so weak that it fails to capture the natural complexity of object relations in the domain of interest (we have, for instance, established above that simple propositional logic is not capable of capturing the syntactic requirements of generalised spatial relations ('within', 'above', 'behind' etc), but may yet suitable for certain limited forms of conceptual inference - see, for instance, the 'Talking Heads' experiment, below).

Given, then, that the modelling of ontological systems in anything other than very limited scenarios will involve the use of at least first order logic, we would appear to have a two-fold learning requirement; a syntactic one (relating to the inference of perceptual interrelationships) and a semantic one (relating to object interrelationships). These will in general be coupled (as in the previous example of percept-action inference). However, the potentially open-ended nature of syntactic inference means that semantic generalisation must occur without a well-defined a priori optimisation landscape. Learning is therefore typically more complex than in familiar stochastic pattern recognition problems, where there exists a fixed feature-space within which observations occur; symbolic ontology learning essentially requires that we learn both the class distributions and the features simultaneously.

We shall hence now turn to a survey of experiments in which spontaneous symbol ontology formation takes place between communicating (or self-communicating) agents in an appropriately limited logical landscape:

### 4.3.2 Spontaneous Language Formation

The study of spontaneous language formation in simulated agents gains its philosophical imperative in consequence of the symbol grounding problem, first enunciated by Harnad [50]. Harnad’s thesis, in distinction to purely structuralist accounts of language, attempts to demand a semantic interpretation of formal symbol systems that transcends the (merely syntactic) interrelationships available to the symbolic manipulation system in question. The problem, as Harnad sees it, is analogous to the learning of non-native languages in humans; it is much more meaningful when attempted in situ amongst other speakers of the language, than when learned from a dictionary of that language alone (as the Structuralists could feasibly envisage occurring).

Harnad consequently proposes two forms of symbolic grounding in particular; 'iconic representations', which are effectively equivalent to class perceptual medians, and 'categorical representations', which consist of both learned and a priori feature invariants. Higher-order symbolic representations are then grounded in hierarchical combinations of these fundamental symbols, so that at no stage is the relationship between generated symbols dictated by the symbol-producing agent alone.

Steels [129], [130] gives perhaps the paradigmatic illustration of the importance of seman-
tic grounding to the formation of language systems in his 'Talking Heads Experiment'. His motivation in these endeavours is to demonstrate that 'communication through language is the main driving force in bootstrapping the representational capacities of intelligent agents'. Language and meaning are consequently co-eval in this scenario; symbolic syntax arises at the same time as semantics.

The talking heads experiment hence consists in a pair of robotic agents equipped with a video camera and a set of predetermined low-level feature descriptors. One agent is initially designated the 'speaker', and the other the 'hearer'. The agents occupy an environment in which regular two dimensional objects of various colours are distributed at random (for instance red squares, blue triangles etc). The designated speaker then chooses one item at random from this common context and attempts to describe it using its own internal lexicon (which it cannot simply assume is shared by the hearer). The hearer must then guess the correct item and point at it, failure to do so requiring the hearer to update its internal lexicon by generating a new word definition that successfully disambiguates the indicated item. The role of hearer and listener are then exchanged over a series of language games in order that an objective world description be finally obtained by both agents (as opposed to the identical, but speaker-subjective one that would arise if the roles of speaker and hearer were fixed). Word definitions are thus characterised in terms of a priori feature descriptors of a visual nature; for instance, colour, horizontal object positions, vertical object positions etc.

As an example of a typical word-game, consider an experimental context in which two objects $A$ and $B$, a red triangle located at the top of the field of view and a blue square located at the bottom of the field of view, are the respective objects of interest.

These might be disambiguated by words descriptors of the form:

$A$: vertical—position $> 0.5$ (positions scaled to $[0 : 1]$)

$B$: vertical—position $< 0.5$

Or, equivalently, by word descriptors of the form:

$A$: red

$B$: blue

(Decision trees can be used to implement more complex designations.)

There is hence no unambiguously 'correct' object ontology in this scenario, and consequently no ground truth perceptual space accessible to the agents. If these two alternative sets of lexical designations were allocated to the speaker and hearer, respectively, it would consequently only be within a fresh context that the discrepancy in description would come to light.

For instance, only if a third blue object were introduced and located towards the bottom of the field of view, would the speaker be required to learn to distinguish the concept of colour as distinct perceptual category (though it always inherently had the latent capacity to do so), in order to distinguish every object employed within the word-game. Equally, the hearer would need to evolve word descriptions that incorporated spatial considerations only in order to distinguish all three objects within the extended

---

10This perhaps correlates with the neonatal synaesthesia hypothesis [51].
scenario.

Steels' achievement is consequently in demonstrating that, within this particular scenario, lexical convergence between speaker and hearer occurs remarkably rapidly. Moreover, provided that there exists a sufficient richness in the range of object scenarios, the talking heads experiment demonstrates that this convergence is objective (in the sense that the final word distinctions correspond to our ground truth descriptions in terms of the a priori features).

This result is consequently consistent with the hypothesis that 'third-person' cognitive modelling lies at the origin of symbol/ontology formation. The objectivity (or subject-independence) of the final convergence of the word designations hence comes about because language conjectures are projected by the speaker back into the environment for validation on the assumption of the presence of a hearer with a linguistic capability similar (in a priori terms) to it own; the modelling of perceptual agency by the agent is thus implicit in the underlying experimental scenario.

In philosophical terms, the talking heads experiment embodies the Wittgensteinian (cf [150]) view of language as 'a word game' in which agents invent words and meanings during their interactions, and opposes the Quinean view that sees language as a series of inductive abstractions of the perceptual correlations between word and object (cf Quine [110]).

Given that the presence of the other agent in this experiment is assumed a priori, we may wonder how far this requirement can be relaxed in language-formation experiments that take place within the simulated agent domain. Viezzer [140] argues that the symbol grounding problem can only really be solved by modelling both the agent's world (at the perception/action) level and the agent's modelling of the world in order to permitting a genuine ontological updating by the agent. This is then in explicit and intentional contrast to the spontaneously arising ontology of the talking-head experiment, which Viezzer claims lacks the semantic updating inherent in word meaning renegotiation (that is, an updating of the ontology on the basis of the agent's model of the agent/world interaction). Implementation of this notion (which would constitute an automated instantiation of the hermeneutic arc, as opposed to the hermeneutic circle of Steel's talking heads), however, remains for future experimental work.

Other attempts to model the emergence of language in embodied cognitive agents utilising a bootstrapping mechanism have typically operated in a more obviously verbal domain. For instance, Narayanan [96] has developed a computational model that takes the cognitive-semantic view proposed by, for example, Lakoff [69], that textual meaning derives from a metaphoric extension of agent causality (so, for instance, textual references to purely abstract entities like the United Nations are envisaged as acting from a spatial position, and being capable of deploying motive forces). Hence, the interpretation of meaning within this context arises from sentence-by-sentence projections of our sensorimotor complex (equivalent to a percept/action mapping) into the domain of discourse, such that our finite memories and goal-driven behaviours are explicitly taken into account.

Thus (to take a simplified example), the newspaper headline 'Economy Falls Into A Hole' describes a purely abstract state of affairs within metaphoric terms. A simple word-by-word mapping of textual definitions in to the semantic space would be incapable of determining the meaning of this sentence; even a contextually sensitive parsing would
be of little use in capturing the full nuance of the sentence’s meaning. The explicit projection of an embodied cognitive agent’s perceptual-action complex onto the noun term ‘Economy’, however, gives rise to a new layer of meaning, within which the sentence would imply both a degree of negligence (we, as agents, falling into a hole would not be looking where we were going), and also the undesirability of this state of affairs and the consequent desire to rectify it (the goal-centricity of our agency would require our immediate exit from the hole in order to continue with our plans). Hence the full meaning of the sentence is only uncovered by the relativising of our individual experience into the objectivity of the noun structure. Lakoff’s argument is that all linguistic constructions operate in this manner (though usually less obviously), with meaning supplied by the agent’s implicit embodiment in the sentence referent.

Narayanan’s computational model of this notion therefore involves constructing analogues of high-level human sensorimotor control that embody sequential, conditional, hierarchical and concurrent actions within a generalisation of automata theory that utilises high-level Petri net variants. A temporally extended probabilistic belief network representing linguistic input and world knowledge is then constrained to a concrete form by mapping it onto this ‘experiential’ model in order to determine the unwritten aspects of the discourse (such as for instance, the implicit desire to ‘get out’ of the economic ‘hole’).

The method is thus profoundly context sensitive by virtue of the explicit modelling of embodied cognition, effectively introducing a finite state machine (representing the agent’s internal goal-states) into the textual interpretation. In essence, Narayanan is employing embodied cognition to provide a semantics to complement the syntax of the language processing.

The methodology adopted by Narayanan can be considered to qualify as a weak form of cognitive bootstrapping, in that it permits agent sensorimotor hypotheses to be overridden by the receipt of new target domain knowledge (which is in turn interpreted in casual agent terms); it does not, however, permit learning of new sensorimotor agent structure from the linguistic target domain, which would be required to complete the definition.

4.4 Summary of Division

We have, in this division, investigated the notion of cognitive bootstrapping from twin perspectives; the connectionist and the symbolic schools of cognitive science, and established its potential consistency with both. We then turned to the issue of symbolic bootstrapping from the cognitive linguistic perspective, and established that communication, whether self-communication, or between agents, is a sufficient condition for the generation of objective symbol ontologies in perceptually self-updating agents. A number of experimental realisations of this notion were demonstrated, in particular Luc Steels’ ‘talking heads’ experiment, which additionally served to illustrate the centrality of communication in overcoming the potential for symbolic bootstrapping to form ungrounded and unrepresentative perceptual categories.

From this empirical outline of the basis for cognitive bootstrapping, we shall now turn, in the next survey division, to an overview of the mathematical underpinning of the concept.
Part IV

5 Foundational Context for Cognitive Bootstrapping

5.1 Formalism Verses Platonism

Of the various foundational philosophies of mathematics that have been proposed, two distinct strands are apparent; the Formalist and the Platonist, the former having arisen as a reaction to the latter. Platonism is the classical view of the ontological status of mathematical entities, and assumes that they are absolutely true and self-subsisting; the objects of perception are, in contrast, transient, unreal and legitimated only to the extent that they embody the abstract truths of mathematics. Mathematical entities are thus absolutely real, having an existence that is independent of cognition.

Implicit in this assumption is the notion that mathematical truths could not be other than what they are. In modern, proof-theoretic terms, this is equivalent to arguing that the logical axioms of mathematics are non-arbitrary.

Formalism, in contrast, assumes no prior existence of mathematical entities at all; mathematics is simply a set of codified schemes for manipulating symbols. As such it is ontologically neutral, making no metaphysical claims beyond its own formal patterning.

From this notion arose the 'Hilbert Programme', the attempt to place all of mathematics on a secure footing by demonstrating its consistency within a formal system (showing, for instance, that it not possible to prove both a statement and contrary within the chosen mathematical axioms). Russell and Whitehead's Principia Mathematica [117] presented the most significant attempt to achieve this.

Formalism was, in the 1930s, to run into two major philosophical difficulties, however. The first was in the continuing success of mathematical physics; many of the latest developments in mathematics such as non-Euclidean geometry and group theory were finding direct application in physics, strongly suggesting that it was anything other than an ontologically neutral subject.

More damagingly, however, was the comprehensive undermining of the objectives of the Hilbert Programme by Gödel's theorem [46], which demonstrated that for any formal axiomatic system, it is always possible to find a statement which is true, but which cannot be formally proved so within the system: it must, in short, constitute a new axiom. The axiom system, if consistent, is thus infinite, and if finite, inconsistent.

Formalism, it thus seems, must fail both in its assumptions of ontological neutrality as well as in its claim to bounded systematicity. This failure was to leave Constructivism as the predominant non-metaphysical alternative to Platonism.

5.2 The Middle Ground of Constructivism

Constructivism (or Intuitionism), as a conceptual reaction to the notion of Platonist mathematics, contends that mathematical objects exist only is so far as they can be constructed. Truth statements applied to entities of infinite extent are thus not permissible unless they can be constructed (the problem within formalism that gave rise to
Gödel’s theorem). The truths of mathematics are thus not empty of content (as they are in Formalism) and not abstracted from human experience (as in Platonism); they are primarily experienced by an thinking agent. The concept of truth is hence modified to that of justification.

Consequently, the most characteristic logical expression of Constructivism is that it rejects the law of the excluded middle, \( (X \lor \neg X) \), but retains the law of contradiction, \( \neg X \rightarrow (X \rightarrow Y) \) (nothing is both \( X \) and \( \neg X \)). In essence, to exert the existence of a thing is to have obtained a proof of it. As such it generalised standard Boolean logic to intuitionistic logic by employing the Heyting algebras (eg [11]). Here, the standard truth values 0, 1 are extended to become partially ordered sets (indicating different stages of truth), allowing a direct isomorphism between mathematical proofs and computer programs to be constructed (the Curry-Howard correspondence). Hence, the notion of mathematical truth gains a temporal aspect consistent with its interpretation in terms of embodied agency.

Several authors have sought to make explicit this connection between mathematically constructed truth and the cognitive capabilities of humans, examples being [145], [124]. Perhaps the strongest statement of the notion that mathematics derives from specifically embodied cognition is given in 'Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being.' [71] in which the authors seek to correlate mathematical structure with metaphors of human agency, finding underlying processes such as object collection, object construction and linear measurement, from which they proceed to derive much of classical mathematics.

These cognitive models of the attainment of mathematical truth are hence resonant of the embodied aspect of cognitive bootstrapping, but do not yet fully encompass its notion of autonomous perceptual updating. For this we need to consider the relationship between axiom and proof more closely:

### 5.3 Cognitive Bootstrapping and Constructivism

Chaitin [21], takes the constructivist idea to the broadest conclusion, viewing not merely mathematical proofs, but also the axioms of mathematics as being, rather than simply assumed a priori, instead derived from our experience, persisting cognitively as conveniently compact representations of our mathematical experience. Using the notion of algorithmic compressibility (which is given a formal treatment in the next section), he demonstrates that it is not possible in general (in the sense of its being incomputable) to establish the smallest possible axiom scheme capable of compactly representing any given mathematical agent’s perceptual history (a mathematical agent being one that proceeds from axiom to proof by arbitrary theorem construction). It is possible, however, to gain partial improvement in the compactness of the mathematical agent’s proof-theoretic experience by adopting differing axiom schema in order to facilitate future proof construction (for instance, by adopting some unprovable but plausible conjecture within the current axiom scheme as a novel axiom, or by adopting as novel axioms the simple results that sometimes arise from complex proof-constructs, such as Fermat’s last theorem).

Given that this implies a lack of absolute validation for any proposed axiomatisation of mathematical truth, Chaitin evolves an ‘experimentalist’ view of mathematics, in which
mathematicians become agents searching (necessarily non-exhaustively) a space of possible truths, happening upon ‘interesting’ axiom systems (like those underlying complex numbers) at random. They are in no way capable of exhausting the possibilities in advance by application of any formal system (such as that envisaged the Hilbert program), and must consequently update their axiom system on the basis of serendipitous discovery and conjecture.

Mathematics as a whole is thus, in Chaitin’s model, an exercise in cognitive bootstrapping; we postulate an initial axiom system (such as those underlying integer arithmetic), and use this as an interpretative basis upon which to make enquiries about further mathematical truth possibilities (for instance, real-number arithmetic), updating our axiom system as we go, and overlooking vast swathes of alternative possibilities (which are necessarily non-encompassable), basing our selections on non-formalisable notions like ‘utility’ or ‘elegance’. Mathematics thus progresses by accident, and eventually arrives convergently on a reasonably stable set of axioms (such as the Peano axioms of set-theory) that have the required, non-formalisable properties.

This view of observer-centred axiomatic contingency is further underlined by the notion [134] that, while the total set of axiom possibilities available to mathematicians may be infinite, their total information content is zero (by extension of the principle that the set of all possible bit-strings of length $n$ can be defined by a (meta)-string of length only $log_2 n$, but a single bit-string would require all $n$ bits). The ensemble of all axioms is thus, in a sense, equivalent to the absence of any axioms. Favouring this notion of the informationless totality of axiom possibilities on the grounds that it constitutes the most general application of the principle of Occam’s Razor to the question of the final ground of mathematical truth would therefore suggest that there is no possible ‘external’ view or final meta-level characterisation of the axiom system from which one could specify a non-observer-relative, absolute axiomatisation. The only thing that acts to constrain our range of possible axiomatisation is then the anthropic self-selection principle, the notion that our mathematical findings must be consistent with our existence as mathematical agents. This implies that the prior axioms that define the scope of possible mathematical theorising in advance (ie possible mathematical experience) are themselves updated purely on the basis of the results of this experience: a cognitive bootstrap, in other words.

While these considerations are largely abstract, and apply only on historical time-scales, there does exist a well-grounded way of approaching the more immediate form of cognitive bootstrapping within a fixed mathematical framework by treating it from a statistical standpoint. Indeed, there exists an established research field concerning the treatment of perception in statistical terms (cf eg [66]) which we shall take as our starting point. The next section will hence look into the application of Bayesian theory to the notion of cognitive bootstrapping.
6 Statistical Context for Cognitive Bootstrapping

6.1 Bayesian Updating and Cognitive Bootstrapping

We shall argue in this section that the statistical technique of Bayesian updating represents the most general framework from which to consider the notion of cognitive bootstrapping. Other approaches to hypothesis updating are, of course, possible, and are perhaps more representative of human cognition: for instance, the modelling of agents that make non-Bayesian assumptions has become a major area of interest in economic theory over the last few years (see for example, [33])\(^{11}\). However, for the present purposes we shall regard all such techniques as either approximations to the ideal of Bayesian updating, or else lacking Bayesian updating’s relative ontological neutrality (particularly in relation to the possibility of determining a universal prior - cf section 6.2.3). Since cognitive bootstrapping is essentially the method by which perceptual ontologies are arrived at, this consideration is of the first importance.

Bayesian methods of inference (eg [5]) differ from classical statistical inference techniques in that probability values are interpreted, not as (asymptotic or finite) trial frequencies, but rather as degrees of belief. Bayesian inference is thus a formalisation of the methodology of science, having the capacity to validate hypotheses or otherwise in relation to experimental data. Bayesian updating is thus the application of Bayesian inference in relation to novel experimental data.

Where this latter method can legitimately be regarded as a precursor of cognitive bootstrapping lies in the fact that the novel experimental data must first be interpreted in terms of the current hypotheses as to the ground-truth model underlying the experimental data, in order that, in an act of reciprocation, their likelihoods can be determined in relation to the totality of the observed data, both novel and pre-existing. The process is thus iteratively updated until model convergence is achieved (assuming that the experimental data is drawn from a singular distribution, and that the range of hypotheses available to the hypothesising agent is sufficient to include, or at least approximate this distribution)\(^{12}\).

Hence, in performing Bayesian updating, we first calculate the probability that a particular hypothesis, \(H\), is true given the experimental data to date, \(X_t\):

\[
P(H|X_t) = \frac{P(H)P(X_t|H)}{\sum_{H_i}P(H_i)P(X_t|H_i)}
\]  

(6)

Here, \(P(H)\) is the degree of prior belief in \(H\) and \(\sum_{H_i}P(X|H_i) = P(X)\) is the degree of prior belief in \(X\); \(P(X|H)\) is hence the extent to which \(H\) can explain \(X\) (or the conditional probability of \(X\) given that \(H\) has occurred in non-subjectivist terms).

When \(H_i\) is parameterised via \(n\) variables \(\theta_j, j = 1\ldots n\), the individual hypotheses, \(H(\tilde{\theta})\) (\(\tilde{\theta} = \theta_1, \theta_2,\ldots \theta_n\)) form the posterior distribution of model parameters:

\(^{11}\)In this research field innate stochastic biases of the human cognitive system, such as realised sample bias and conformation bias, can be demonstrated to have a measurable cumulative effect on economic indicators such as stock prices.

\(^{12}\)Note that we are not, here, making a connection with parametric bootstrapping (eg [57]); the quantity we are ‘bootstrapping’ into existence are the model parameters, not the sample data.
\[ P(\tilde{\theta}|X_t) = \frac{P(\tilde{\theta})P(X_t|\tilde{\theta})}{\int_{\theta} P(\theta)P(X_t|\theta)} \]  

\[ (P(\tilde{\theta}) \text{ is the prior parameter distribution, often assumed to be uniform.}) \]

Critically, even for large variations in the prior parameter probability distribution assumptions, Bayesian observers will tend to find the posterior probabilities asymptotically convergent.

The model parameters are hence \textit{constrained} by the observational data. It is consequently possible to make quantitative predictions in relation to novel data possibilities, \( X \), in the light of these constraints:

\[ P(X|X_t) = \int_{\tilde{\theta}} P(X|\tilde{\theta})P(\tilde{\theta}|X_t) \]  

\[ (\text{actualised data is then appended to the existing data } X_t \text{ to give the next temporal iteration of the experimental data, } X_{t+1}. ) \]

It is also possible in this framework to consider models in which memory resource is finite, in which case only a given, fixed quantity of experimental data can be retained at each iteration. Sometimes the ‘forgetting’ implied by this can even assist the process of posterior convergence, since the impact of poor initial bootstrap priors are quickly removed from the system.

Whichever form is adopted, at each iteration of Bayesian updating our range of plausible world parameterisations becomes, in a sense, the perspective through which we interpret \textit{actual} world data; data has a different \textit{meaning} within the differing parameter-models (as indicated by their differing predictive functionalities), even though this data is itself used to constrain the models. This interpretive aspect is thus more explicit than in conventional Bayesian inference, where model inference in relation to the experimental data takes place in a simultaneous batch.

It is hence in this potentially paradoxical iteration between interpretation and interpretative model inference in which the relevance of Bayesian updating to cognitive-bootstrapping lies, although the parallel with Dilthey’s notion of the hermeneutic circle cannot yet be formally drawn. This possibility only becomes fully realised on considering \textit{agent-based} Bayesian updating (eg [45]), wherein the model conjectures become the basis upon which \textit{experimental testing}, as opposed to data interpretation, proceeds (the local consistency checking of globally-attributed word-meaning hypotheses in the hermeneutic circle of interpretation may be considered a form of experimental testing in so far as it involves the selection of a \textit{particular} hypothesis for actualisation in relation to which the other hypotheses are refined).

Predictive models are hence used in agent-based Bayesian updating to \textit{guide} the data seeking process, as opposed to passively accepting novel data as is the case in the non-agent-based scenario. Thus, for instance, a mobile robot might employ a partially-constrained model of wall orientations within a room in order to navigate towards hypothesised corner-locations so as to provide better constraint information for that same partial hypothesis; uncertainties can thus be made to swiftly converge to zero in an \textit{intentional} fashion. We therefore now turn explicitly to the subset of Bayesian updating known as Bayesian \textit{exploration}.

58
6.1.1 Bayesian Exploration

Bayesian exploration differs from conventional Bayesian updating in being active, rather than passive. Hence, rather than a Bayesian agent being limited to interpreting externally received data, Bayesian exploration involves the active seeking of data with which to confirm or deny its environmental hypotheses. In this way, the method constitutes a constrained realisation of the concept of cognitive bootstrapping, since the inferred higher level perceptual hypotheses (for instance environmental maps\textsuperscript{13}) by which the agent’s data-seeking actions are governed derive from low-level a priori percept-data (for instance radar distance measurements) that themselves remain exempt from the perceptual updating process. In this way the potential chicken-and-the-egg paradox is averted, in which the agent must interpret low-level sensory data in terms of its high level perceptual inferences (the agent ‘positions’ novel incoming sensory data relative to the currently favoured environmental map by which it is manoeuvring), and yet must also build this high-level perception from that same low-level data.

A very general mathematical realisation of this idea in terms of autonomous mobile agency comes from [125]. Here, Sim asserts that when given a particular sensor reading \( z \), and a ‘pose’ \( q \) (defined as the global parameter vector defining the state of the agent), the likelihood of a particular ‘map’ \( m \) being correct, given the agent’s current state, is given by the Bayesian rule:

\[
P(m|q, z) = \frac{P(z|q, m)P(m|q)}{P(z|q)} \tag{9}
\]

The map is thus understood as a predictor for agent observations given a particular pose.

Exploration, in this context, is considered the attempt to establish the pose \( q \) that maximises the expected entropy reduction in the distribution \( P(m|q, z) \).

This change in entropy is given in terms of the cross entropy:

\[
G = - \int_{m \in M} P(m|q, z) \log \frac{P(m)}{P(m|q, z)} \, dm \tag{10}
\]

The expectation of the change of entropy (the term whose magnitude is to be maximised) is thus:

\[
E(G) = - \int_{z \in Z} G \, dz' \tag{11}
\]

\textsuperscript{13}An environmental map would be regarded as an object model when not considering Bayesian exploration. However, it is here employed as the both the final object model and the sensory ordering concept through which novel data is interpreted during the exploratory phase, thereby qualifying as a perceptual categorisation for half of the iterative cycle.

In fully unconstrained cognitive bootstrapping, however, we would expect to have the potential for inference of novel perceptual categories with the capability of delineating the possibility of particular objects, but which could not, in themselves, be equated with any particular object so conceived. Percept categories thus, in general, exist at a meta-level with respect to perceived objects: however, note that in a hierarchically open-ended cognitive systems percept-categorisations can themselves become objects as further meta-categorisation are conceived.
Parameterisation of $m$ is not always obvious; the author, however, conjectures that $m$ can, at worst, be parameterised as the total set of observations: $m \in \mathbb{Z} \times \mathbb{Z} \times \ldots \times \mathbb{Z}$.

The analysis thus far assumes that pose is known; of course this is not the case for the autonomous, bootstrapping agent. Rather, the pose must be inferred from the sensory data and the map estimates. That is, the map (which serves to predict agent observations given a particular pose) must itself be used to determine the pose, the tautological perceptual problem we have identified as necessitating the cognitive bootstrap mechanism.

The only way to resolve this problem is thus to invoke the a priori certainty of the motor-space of actions, $a$. The agent’s goal must therefore be to select an action $a$ that maximises the expectation $E(G)$ of the whole configuration space, $C$:

$$E(E(G)|a) = - \int_{q \in C} E(G) \, dq$$

(this need not necessarily be as exhaustive as indicated; posteriors often need only be partially sampled: this is Monte Carlo updating [73], an approach that finds many applications within the literature.)

This latter equation consequently represents, in mathematical terms, a relatively generalised conception of cognitive bootstrapping for Bayesian agents, given the implicit perceptual restrictions; in fact, the information-theoretically optimal form of the non-open-ended variety of cognitive bootstrapping for agents occupying a stochastically-determined environment.

The most general forms of Bayesian-updating, however, are those in which we do not limit the range of parametric possibilities in advance. Hence, rather than regarding unsampled parameters as non-existent, we can instead regard them as not yet determined, allowing for the possibility of emergent models of cognition. Bayesian exploration can hence be made to simultaneously explore the perceptual space as well as the environmental space, becoming, in the process, entirely autonomous. This type of approach, in which the scope of discoverable truth is not limited in advance, in hence resonant with the mathematically constructivist notion of intuitionistic logic [17], in which the law of the excluded middle is eliminated in favour of temporally-dependant truth values (‘true’ being reinterpreted as ‘not yet falsified’, as befits a finite, embodied, and consequently non-omniscient logical agent).

One such practical framework for formalising this approach mathematically is that of the Dirichlet processes.

### 6.1.2 Dirichlet Processes: A Framework for Non-Parametric Bayesian Updating

Dirichlet processes arise in the study of nonparametric Bayesian theory [34] as a means of determining how new model components are generated in response to novel experimental data. These components are hence generated from an infinite-dimensional distributional meta-parameter, $G_0$, with a frequency governed by a second meta-parameter, $\alpha$. $\alpha$ can, in general, be determined directly from the experimental data via maximum likelihood estimation: $G_0$ has usually to be assumed, though work exists (eg [152, 81])
to suggest that this too may be estimated.

Typically, Dirichlet processes are employed in the context of mixture models (see, for example, [112], whose terminology we follow), for which \( \alpha \) determines the rate of formation of mixture components. Mixture models are defined as having additive components of identical functional form, \( f \), but with differing internal parameters, \( \theta_i \) (which are hence drawn from \( G_0 \)): they also differ in their external multipliers, \( \pi_i \). In experimental scenarios in which Dirichlet processes are assumed to be operating, we thus typically have \( N \) individual measurements, \( y_i \), drawn from an unknown number of mixtures \( k \). The probability of each individual measurement is, for a model of fixed \( k \):

\[
P(y_i) = \sum_{j=1}^{k} \pi_j f(y_i|\theta_j)
\]

(Each measurement is assumed to be generated by a single component).

A significant problem is thus to determine the number of mixture components \( k \) to consider in relation to the observed data.

To begin to address this, we first consider the Dirichlet distribution. The Dirichlet distribution is derived from the Polya Urn metaphor, in which a bag of \( \alpha \) balls have colours labelled \( j \) with initial frequency \( m_j \). Balls are then drawn at random and replaced by two balls of the same colour. It is straightforward to show that, for a given observational history \( y_{1:N} \), the probability of finding that the next observation, \( y_{N+1} \), is of colour \( j \) is:

\[
P(y_{N+1}|y_{1:N}) = \frac{\alpha m_j + \sum_{i=1}^{N} \delta(y_i - j)}{\alpha + N}
\]

(\( \delta \) is the Kronecker delta)

The Dirichlet process is a continuous version of this formula in which \( k \to \infty \) and \( M \) becomes continuous, giving the Blackwell-MacQueen formula:

\[
P(y_{N+1}|y_{1:N}) = \left\{ \begin{array}{ll} \frac{1}{\alpha + N} \delta(y_i - j) & \exists l \leq N : y_l = j \\ \frac{1}{\alpha + N} M(j) & y_l \neq j, \forall 1 < l < N \end{array} \right.
\]

Hence, in the Dirichlet process mixture model, for each measurement \( y_i \), a corresponding \( \theta_i \) is drawn from \( G_0(\theta) \), which substitutes for \( M \) (the continuous base measure in the Blackwell-MacQueen formulation):

\[
P(\theta_i = \theta | \theta_1, \theta_2, \ldots, \theta_{i-1}, \alpha, G_0) = \left\{ \begin{array}{ll} \frac{1}{\alpha + i - 1} \sum_{j=0}^{i-1} \delta(\theta - \theta_j) & \exists j \leq i : \theta_j = \theta \\ \frac{1}{\alpha + \gamma} G_0 & \forall j < i, \theta_j \neq \theta \end{array} \right.
\]

Combining this with Bayes theorem, we thus have a formula for determining the posterior probability distribution of \( \theta_i \) given a particular data instance \( y_i \):

\[
P(\theta_i = \theta | \theta_1, \theta_2, \ldots, \theta_{i-1}, \alpha, G_0, y_i) \propto f(y_i|\theta_i)P(\theta_i = \theta | \theta_1, \theta_2, \ldots, \theta_{i-1}, \alpha, G_0)
\]
We can thus begin to appreciate how the Dirichlet process can be used to constrain unsampled parameters in potentially infinite parameter-space models using only the currently detected data. Dirichlet process are a hence one practical framework for performing open-ended, non-parametric Bayesian updating of the type required for emergent cognition, finding agent-based application in, for instance, environmental topology mapping within mobile robots in [113].

Following this practical indication of how it is possible to limit the constraints inherent in parametric Bayesian exploration to allow for truly emergent perceptual possibilities (so that, for instance, a mobile robotic-agent could determine an appropriate perceptual/motor scaling parameter to employ on the fly), we should now like to look at the possibility of absolutely unconstrained approaches to Bayesian updating, wherein no a priori mathematical form is imposed in advance.

6.2 Algorithmic-Information Theoretic Frameworks for Fully Unconstrained Bayesian Updating

In order to arrive at the most general form of Bayesian updating, we shall first need to explore the notion of Kolmogorov complexity, an algorithmic generalisation of the concept of informational entropy. In consequence of this conceptual origin, Kolmogorov complexity is often referred to as 'algorithmic entropy' in the literature.

Before defining the quantity formally, a few terminological determinations need to be given: \( l(x) \) shall denote the number of bits in a given bitstring, \( x \). The equational form \( T(p) = x \) will be understood to state that the Turing machine, \( T \), with input, \( p \), computes \( x \) and then halts.

\( U^c \) is then a defined as a universal Turing machine if there exists a constant string (a compiler), \( p_T \), such that for all Turing machines, \( T \in C \), \( U^c(p_T) = x \).

The Kolmogorov complexity \( K(x) \) of a bitstring \( x \) can now be defined, for a fixed \( U \), as:

\[
K_U(x) = \min_p \{ l(p) : U(p) = x \}
\]

(18)

The Kolmogorov complexity is hence the minimum-lengthed Turing machine that computes \( x \) (with respect to \( U \)).

We can hence straightaway infer from this that the majority of bitstrings, \( x \), are random, which is to say, algorithmically incompressible, since, for a given \( n \), there are a total of \( 2^n \) programs with \( \leq n \) bits, but \( >> 2^n \) bitstrings with \( > n \) bits.

A secondary consequence of this definition of algorithmic complexity arises as a consequence of the notion of a compiler; we find that for any two universal Turing machines, \( U_1 \) and \( U_2 \) the relation \( K_{U_1}(x) = K_{U_2}(x) + O(1) \) holds. We do not therefore need to worry about the particular specification of \( U \), and the Kolmogorov complexity of a particular bitstring \( x \) achieves a near-universal form.

However, as it stands, this definition of algorithmic complexity is incomputable for any given \( U \) as a consequence of the Halting problem. The notion is therefore necessarily an abstract one; this does not, though, prevent us from utilising the concept in order to refine our notion of Bayesian updating.
6.2.1 The Bayesian Prior Problem

Perhaps the most controversial aspect of Bayesian theories concerns the specification of the distribution of priors, $P(H)$. In a constrained environment, however, we are often justified in assuming uniform prior distributions over particular parameters such as the kernel standard deviation. This is particularly so if Bayesian updating tends rapidly towards a convergent solution independent of the prior distribution.

However, a free-form, open-ended cognitive agent capable of learning complex and novel models cannot freely make this type of a priori assumption. We have, in the previous section, looked at how model order parameters might be made open-ended; however this technique necessarily assumes a fixed kernel form. What we would ultimately like to achieve is a Bayesian updating system that, in addition to making no prior parameter distribution assumptions, also makes no prior model assumptions, thereby allowing a completely general form of $P(H)$ to emerge.

One way to achieve this is to invoke the minimum description length (MDL) principle:

6.2.2 Bayes Theory and the Minimum Description Length Principle

The minimum description length principle asserts that the most plausible hypothesis for explaining a given set of empirical data, $X$, is the one that minimises the sum of the hypothesis’s description length in bits ($i(H)$) and the data’s description length in bits when that data is encoded by that hypothesis, ($i_H(X)$). In other words, the MDL principle asserts Occam’s Razor in terms of the total information content of both the model and the model-interpreted data.

If we formalise this more exactly in information-theoretic terms, the description length of $H$ is given by the quantity $-\log P(H)$, the description length of $X$ in terms of (ie given) the hypothesis $H$ is given by the quantity $-\log P(X|H)$, and the description length of hypothesis $H$ in terms of (given) the data $X$ is given by the quantity $-\log P(H|X)$. Therefore, the MDL principle states that we wish to obtain the hypothesis $H$ that minimises the quantity:

$$-\log P(H|X) = -\log P(X|H) - \log P(H)$$

which is equivalent to minimising:

$$-\log P(H|X) = -\log P(X|H) - \log P(H) + C$$

for $C$ an arbitrary constant.

Letting $-C$ equate to the description length of the data $-\log P(X)$ (which is a constant in this scenario), we obtain:

$$-\log P(H|X) = -\log P(X|H) - \log P(H) + \log P(X)$$

Exponentiation of this quantity suggests that the MDL principle requires that we find the $H$ that maximises:
\[ P(H|X) = \frac{P(X|H)P(H)}{P(X)} \]  

This, however, is \textit{exactly} equivalent to Bayes theorem. Hence, by inversion, the Bayes approach is simply a restatement of the minimum description principle in probabilistic terms.

Equipped with this reformulation of Bayes theorem, as well our earlier understanding of Kolmogorov complexity, it now becomes possible to address the Bayesian prior problem and propose a form for \( P(H) \).

6.2.3 A Universal Prior for Bayesian Updating

Following the description of Kolmogorov complexity, we are now in a position to define the prior probability of a single piece of data represented by the bitstring \( X \) as being \( P_{U_1}(X) \), the probability of randomly \textit{guessing} the program \( p \) that computes it on \( U_1 \) (we shall assume a uniform distribution over the binary digits \([0,1] \)). This is given, for equal probability random binary digits, by the Solomonoff-Levin distribution:

\[ P_{U_1}(X) = \sum_p \left\{ p : U_1(p) = X \frac{1}{2} \right\} \]  

From the compiler theorem we have that adopting an alternative universal Turing machine, \( U_2 \), modifies the above to:

\[ P_{U_1}(X) = \sum_p \left\{ p : U_2(prp) = X \frac{1}{2} \right\} \]  

however, by definition;

\[ P_{U_2}(X) = \sum_p \left\{ p : U_2(prp) = X \frac{1}{2} \right\} \]  

and hence;

\[ P_{U_2}(X) = \frac{1}{2} P_{U_1}(X) \]  

That is, the two probabilities differ by only a constant factor (the probability of correctly guessing the compiler, in fact). Probability ratios for individual bitstrings are therefore independent of the universal Turing machine chosen, and the prior probability distribution over the array of possible bitstrings, \( P_U \) is hence \textit{universal}.

It is, moreover, possible to show \([72]\) that \( P_U(X) = O \left( 2^{-K_U(X)} \right) \), or that the probability of guessing at random \textit{any} computational model of \( X \) is essentially equal to guessing its \textit{simplest}.

We can hence find an algorithmic justification of Occam’s Razor by looking at an extension of the data sequence \( X, X' \) (such that \( X \in X' \));
\[ P_U(X'|X) = \frac{P(X')}{P(X)} = O \left( \frac{2^{-K_U(X')}}{2^{-K_U(X)}} \right) = O \left( \frac{1}{2^{K_U(X') - K_U(X)}} \right) \]  

In other words, the most probable observational continuation of an observed sequence \( X \) is the one with the smallest additional Kolmogorov complexity; the simplest theory to fit any set of observations is therefore the more likely.

The above strongly suggests the application of algorithmic theory for achieving the most general form of Bayesian updating/cognitive bootstrapping. However, we are at present limited by the incomputability of \( K(X) \) (although some authors, for instance, [101], would argue that these barriers are not actually applicable to truly cognitive entities).

We therefore turn to a consideration of how this difficulty might be overcome practically:

### 6.2.4 Practical Implementation of Algorithmic Bayesian Updating

The incomputability of \( K(X) \) manifests itself as the failure to halt of \( U_c(p) \), a situation that invariably arises as soon as \( p \) becomes equal to the index of any program that incorporates \( U_c \) itself along with another output. In this case we have the impossible situation of \( U_c \) attempting to compute its own final output plus some additional output value; a situation that cannot meaningfully terminate in the result state other than for certain inconsistent axiom sets.

We have therefore always to work with a finite temporal resource in order to practically utilise Kolmogorov complexity. This is captured by the notion of Levin complexity [72]:

\[ K_l(X) = \min_p l(p) + \log T^K_{p} : U(p) = X \]  

Here \( T^K_{p} \) is the number of iterations or computational steps that the Universal Turing machine of index \( p \) must undergo before arriving at \( X \) (note that this is not now the output of a halted program, necessarily).

An implementable form of the conditional probability of a sequence \( X \) over the universal prior, \( P(X|H) \) can now be determined via Levin’s universal search algorithm, which provably achieves the optimal performance for a fixed temporal budget.

Levin’s universal search algorithm thus generates all \( p \) in order of increasing \( K_l \). Hence the universal search implements a series of phases, \( i, i = 1, \ldots k \) which run in lexicographic order all self-delimiting programs \( p \) of length less than \( k \) for \( 2^i2^{-l(p)} \) iterations, where \( k \) is the first such that the condition \( U(k) = X \) holds.

A practical implementation of the above is given by [62] in the context of neural-network parameterisation. The method is of potentially very much more general utility, of course; indeed it can be reasonably argued that the Universal Search method is the most general model inversion (and hence inferential learning) methodology.

For this general complexity schema to be applicable to cognitive bootstrapping, as opposed to merely object model regression, we need to revert to our former terminology and write \( p = prp' \), that is, we need to decompose the index \( p \) into its separate components: \( prp' \), the index of the Turing machine \( T' \) emulated by the Universal Turing machine \( U \), and \( p' \), the input to this Turing machine. Doing so allows us to distinguish
of $T'$ as the perceptual mechanism by which the a priori observational sequence $X$ is reinterpreted as the perceptual datum, $p'$. Hence Levin complexity, in minimizing $p$, thus serves to simultaneously update both the perceptual mechanism and, by implication, the perceptual datum (which is to say, the object model).

If we hence explicitly incorporate an experimental-validation phase, so that $p'$ corresponds, instead, to the alternating sequence $[p'_1, a'_1, p'_2, a'_2, \ldots]$, where the $a_n$ are the hypothesis-testing exploratory actions giving rise to perceptual observations $p_n$, then an increase in Levin complexity that is log-linear with time would indicate that the perceptual model has converged.

More generally, since the perceptual updating aspect of cognitive bootstrapping is essentially the act of compressing the interpretative framework of perception with respect to the perceptions that this compression gives rise to, the retention of a high-level perception-action link that is hierarchically derived from the a priori perception-action link allows the compression to be tested for its generality. Compressibility hence serves only as the basis for the perceptual updating process; it is actions (in terms of the percepts that they give rise to) that serve as validation criterion for the perceptual updating.

6.3 Summary of Division

We have, in this part of the survey, looked both at the mathematical underpinnings of the concept of cognitive bootstrapping, and conversely, within the foundational section, at the cognitive bootstrap underpinning for the concept of mathematics.

We hence have a two-tier notion of cognitive bootstrapping: in so far as we take the axiom schema of mathematics to be a given, we have proposed a universal model of cognitive bootstrapping in terms of agent-based Bayesian updating carried out within an algorithmic-information theoretic framework, finding an ideal mathematical form in the notion of Bayesian exploration: on the other hand, in so far as we can regard the axiom schema itself as subject to empirical selection (in Chaitin’s model), we can regard the exercise of mathematics as itself the form of a cognitive bootstrap. We regard the most immediately practically-applicable description of cognitive bootstrapping for embodied artificial intelligence (robotics) as that expressed by equation 12, which captures the notion of the non-reinterpretability of a priori perception/action categories that must, according to our Kantian analysis, underpin perceptual updating within autonomous agents.

Similarly, and in keeping with the document’s structural progression from abstraction towards application, for the next part of the survey (in which we explore the computational perspective on cognitive bootstrapping), we shall hence consider the axiom schema to be exactly the standard Zermelo-Fraenkel axioms of set theory, and further that the cognitive bootstrapping is of a non-universal kind (none of the mechanisms surveyed employ the computable Levin search form of Kolmogorov inference within their bootstrapping methods). Hence, it remains for the most generally implementable form of cognitive bootstrapping outlined in this section to be constructed.
Part V

6.4 A Note on Bootstrapping Methods in General Computational Science

Bootstrap methods are part of the common currency of computer science, particularly with respect to issues of language origination. Here, the paradox to which the bootstrapping process is addressed involves the notion of Turing equivalence; once a language has sufficient expressive power to be determined Turing-complete, it is capable of expressing any computation, and therefore of emulating any computational language. In particular, it is capable of emulating itself. Thus, once a minimal subset of a high-level language has attained Turing-complete functionality, it can then be used to extend itself; the language has, in a sense, begun to write itself, or self-compile.

Very often, once this has been achieved, the original Turing-complete subset of the language used to bootstrap the rest of language into existence can be forgotten, or even omitted. In certain very-high level languages there is consequently no trace of the originary language functionality, and hence it appears to have 'pulled itself up by its own bootstraps'.

A similar situation is encountered in the initialisation of computers with complex operating systems; an operating system is required to mediate between hardware and software, however, an operating system must, as software, first be loaded on the hardware (requiring a further interface between hardware and software, and so on ad infinitum). This paradox of infinite regression is resolved by having a small inbuilt (ROM-based) operating system called the bootstrap loader with just enough functionality to achieve the loading task in hardware. Bootstrap loading, in its abbreviated form, hence gives us the term 'booting'.

The familiarity of bootstrap ideas within general computing has meant that they are naturally applied to the problem of autonomous artificial agency. The following sections will therefore outline a representative number of such approaches, in so far as the resulting system of bootstrapping can reasonably be considered cognitive, if only within its domain of operation.

7 Cognitive Bootstrapping Methods in the Field of Artificial Intelligence

Various modalities of cognitive bootstrapping exist within the field of artificial intelligence, but two broad characterisations are, in particular, apparent: the embodied and the disembodied. We have already seen a number of examples of the latter in the cognitive linguistics section of the survey; we should like, in the following section, however, to focus on the specific aspect of situatedness within cognitive bootstrapping. That is, we wish to look at artificial implementations of the cognitive bootstrap that directly mediate between an artificial conceptualising agent and the real world by virtue of the embodiment of the former in the latter.
7.1 Robotics

The engineering field in which embodied cognitive bootstrapping receives its most tangible expression is thus robotics; the study of programmable machine systems. When this programmability extends to the notion of self-programmability, we are concerned with the particular field subset known as autonomous robotics.

When the goal of an autonomous robot is to construct a sensory model of its (presumably previously unexperienced) environment, we are then implicitly in the realm of artificial embodied cognition. While early attempts at robotics typically sought to solve this problem at the sensor level, for instance by using radio markers or range-finding, recent advances in the computer processing power available for real-time computation have allowed robotics to begin to employ cognitive vision methods, for which the sensory input consists of mono-, stereo- or multi-scopic camera feeds.

Environmental modelling in the cognitive vision regime is hence analogous to that exhibited by the mammalian cognitive vision system (particularly when dealing with with stereo and multiscopic camera feeds, for which the primary computational burden is the three-dimensional reconstruction of the environment from planar projections). Typical low-level cognitive tasks thus include edge detection, object segmentation, motion registration, and so on, with potentially ever higher levels of cognitive abstraction possible beyond the immediate base-level vision tasks.

The degree to which environmental cognition can be made fully open-ended is thus a matter of architecture; however, it is necessary, or at least, vastly simplifying, to incorporate a number of a priori constraints on the cognitive reinterpretation process, the general minimum being the presence of a sensory topology that defines the arena in which the autonomous robot is active as a space. Cognitive bootstrapping can thus, in its fuller implementations, be essentially unbounded within the usual Kantian constraints, or else can occur within prescribed limits that go well beyond those required a priori: both of these possibilities are considered in the following sections.

We shall hence set out to review a representative range of cognitive modelling approaches to autonomous robotics, and shall, in keeping with the overall structure of the survey, commence with the most open-ended such approaches (that is, ones in which a priori perceptual prereconceptions are kept to a minimum), moving on to the more limited cognitive bootstrap models, such as SLAM robotics.

7.1.1 Open-Ended Cognitive Bootstrapping in Autonomous Robots

Hierarchical Percept-Action Approaches  Hierarchical percept-action approaches to artificial cognition were first proposed by [15], who employed the term subsumption architecture. This approach was this intended to mimic the connection between sensation and action found in biological agents, and was primarily intended to overcome the philosophical difficulties with the concept of representation associated with classical approaches to artificial cognition. Here, only the external responses of the cognitive agent to particular environmental stimuli are considered, reducing the possibly subjective concept of agent representation to the set of objective actions obtained in response to external stimuli.

The subsumption architecture hence admits of hierarchicality, wherein the higher ar-
chitectural layers control the behaviour of the lower via the mechanism of inhibition, allowing the possibility of open-ended development of the cognitive agent's responses. However, the possibility of abstract symbolic inference in these higher hierarchical levels is not directly considered by Brooks.

Modavil [91] hence proposes a method of bootstrapping progressively higher levels of symbolic representation, up to and including the concept of objects, via the clustering of representations from previous levels within his OPAL architecture (Object Perception and Action Learner). Bootstrap learning thus allows the system to move from egocentric (view-centred) and allocentric (object-centred) sub-symbolic descriptions to symbolic object-based description by ascending a four-fold hierarchy; Individuation, Tracking, Image Description and Categorisation.

Individuation involves the use of occupancy grids to classify individual sensor readings as either static or dynamic. Clusters of dynamic readings are then tracked over time to provide an object model; stable shape models are then constructed from the consistent aspects of the objects so formed. OPAL is thus capable of autonomously discretizing the sensory environment into a static background, the learning robot, and a set of movable objects.

It is then intended that control laws be bound to the object representations so that the robot can define a set of affordances and thus engage in goal-centric activity. These goals can in turn refine the object description so as to make them truly agent-relative.

Granlund [47] provides a still more general architecture for cognitive robotics based on the notion that scene description is not required prior to action. Thus, it is argued that the failure of conventional cognitive architectures is due to the categoric abstraction of objects occurring at an intermediate stage between percept formation and action specification. What is lost in this approach are the contextual modifiers necessary for precise specification of agent action; in short, we gain descriptivity at the expense of intentionality, the latter being relevant only to an embodied agent in a particular context\(^\text{14}\).

Granlund hence proposes a bootstrap mechanism for the initial learning of the embodied system based on a perception-action feedback cycle. Here, in the learning phase of the perception-action mapping, action always precedes percept. Thus, the potentially enormous complexity of the percept domain is limited by considering only those percepts related to actions, which thus occupy a far smaller state-space (an idea of the information-theoretic disparity between these two different types of environmental modelling, the agent-specific and the agent-non-specific, can be found in [97]). In the absence of prior scene-knowledge actions are hence driven by biologically-motivated random exploration impulses (literally random walks in the action state space).

The percept-action mapping can thus be made subject to various optimisation procedures that allow compact representation, and implicitly, therefore, generalisation. The random actions and subsequent compact percept mappings thus amount to an unsupervised training of the architecture. There consequently exists a natural stopping criterion for the random action impulses at the point at which the compact represen-

\(^{14}\text{A counter argument to this type of approach, however, is given in [2], which suggests that agent-based reinforcement learning is more effective when indirect (ie those which build a full agent-independent environmental model) than when direct (ie those which learn directly from the sensorimotor experiences and abstract the environment model).}
tation of the percept-action mapping no longer undergoes significant change (learning having converged). At this point the random action impulses can ascend to a greater level of abstraction and operate on the higher-level percept-action representations that have been generated by the compact generalisation. Of course, these higher level action impulses themselves generate further training data for the largely trained lower levels, allowing for robust and adaptive learning across the whole of the hierarchical structure so formed.\footnote{Thus, to give a concrete human example of this type of hierarchical perception-action mapping, we might initially learn to ride a bicycle by experiencing the way in which the handle bars and pedals react to our actions independently of each other. Only after the concepts of 'steering' and 'pedalling' have become well-established to cognition, having represented the totality of our experience in the two respective areas in a compact and generalised fashion, can we then proceed to learn a high-level concept such as 'cornering', requiring exploration of the space of possible \textit{conjunctions} of the concepts of 'steering' and 'pedalling'.}

These compact representations within the hierarchical percept-action structure are \textit{symbols}, and correspond precisely to the symbols employed in verbal communication. Such communication can hence be considered a low-bandwidth interaction between agents that allows complex actions to be initiated in one agent by another by virtue of the 'unpacking' of the compact representations that takes place as information travels down the percept-action hierarchy from the highest to the lowest levels. Symbolic communication between such agents is hence \textit{always} grounded.

Meaningful communication between agents must also presuppose compact symbolic representations that are \textit{independent} of particular agents' current situations, even though they must be directly \textit{applicable} to it. This amounts to a requirement that communication is \textit{objective}; it must thus refer to object-centric descriptions. Hence a verbal instruction from one agent to another might refer to the placing of one object on top of another; it is then for the individual agents to relate this perspectiveless description to their immediate experience of a perspectival collection of low-level percepts (orientated lighting-gradients, edges etc).

Thus, for a communicative agent, the view-dependant description must co-exist with the object-centred description in a manner that recalls the discussion of the relationship of the transcendental unity of apperception to the \textit{a priori} object concept set out in the Philosophy division of this survey. The difference here is that the object concept is derived from the \textit{a priori} possibility of compact, generalised object-percept mappings, while for Kant the object concept is itself an \textit{a priori} concept: Granlund's \textit{a priori} cognitive generalisation structure is not, in itself, explicitly \textit{geometric}.

Granlund further proposes a specific data representation for the percept-action pairings that naturally allows for these mode separations and generalisations; the \textit{channel representation}, which ensures a fast convergence of the learning phase by virtue of its tendency to identify local subsystems.

The cognitive architecture thus defined is clearly one of cognitive bootstrapping; the inferred higher-level cognitive hypotheses validate themselves in terms of the lower-level hypotheses and vice-versa by virtue of the 'filtering-down' effect wherein exploratory impulses in the conjectural high-level abstracted cognitive categories (for instance, a sensorimotor intention such as driving a car to a particular location) will result in progressively more \textit{contextualised} low-level actions. Hence, the lower-level cognitive categories that perceive the immediate locality in terms of edges and vertices etc, pro-
gressively ground the abstract instruction in the real world and allow for the possibility of empirical validation of the high-level inferences in a way that would be impossible for a purely symbolic or 'top-down' system.

Sun [133], in setting out a foundation for artificial cognitive architectures, similarly argues that human cognition is essentially 'bottom-up' and further, that minimal initial bootstrap models are necessary to avoid over-representational models that may fail to generalise.

Stein [131] also argues that goal-based behaviour in cognitive robots should be considered, not at an abstract symbolic level, but rather at the lowest sensorimotor levels. Hence, in projecting a goal, a robotic agent should utilise exactly the same exploratory and learning processes that it uses to interact with the real world, but instead substitute a 'virtual reality' interface at the very lowest level of the sensors and actuators. This virtual reality is precisely the sensory map formed by the world-model currently hypothesised to be the most likely. 'Cognition', for Stein, is hence simply the imagined sensation and action implicit in tracing out an action path to a particular goal state in the model.

Stein’s MetaToto hence self-trains its higher-level cognitive abilities using only its internal representations (which may be independently learned). Thus, for instance, the MetaToto-styled robotic agent might have an internal percept-action model of a wine-bottle and a wine-glass, and proceed to learn the concept of 'pouring' via this simplified internal model, rather than in the real-world with its attendant complexities.

Generalisation is thus absolutely implicit in such a system; MetaToto is capable of imagining any plausible extrapolation of the currently experienced portion of the world, and training itself to behave appropriately in all of these domains. It can thus construct navigation systems that can cope with any theoretical terrain by building alternative environments using random placements of existing landmarks and exploring these in terms of its own sensorimotor responses.

We might speculate (though Stein does not) that there is a Darwinian justification for this imaginative self-training; a biological agent that tests its action hypotheses in imagination can rule out potentially unsurvivable actions without endangering itself. Imaginative agents are thus more likely to prevail and reproduce than equivalent unreflective agents. In human terms, this principle may also relate to the phenomenon of sleep paralysis, which we shall treat more completely in the discussion and conclusions section.

In the newly emerging, biologically-motivated field of 'cognitive developmental robotics' Asada et al. [3] propose, by analogy with human cognition, that robot design should not fully embody either the nature or the nurture paradigms. That is, robotic agents should neither have their behaviour predetermined by a programmer, nor have it derived via purely environmental learning, since if either of these strategies are fully implicit no new knowledge is possible. In the former case, this is because hard-coded behaviour impedes new interaction possibilities; in the latter case this is a consequence of indiscriminate data collection which lacks any intrinsic meaning without a set of well defined agent capabilities to ground it.

Asada et al's proposed cognitive framework thus acknowledges the necessary a priori structure that must underly any form of cognitive bootstrapping.
Another framework for open-ended cognitive induction in autonomous robots is set out by Nagai et al. [94]. Their central concern, 'Joint attention', is analogous to the object of Steel's 'talking heads' experiment detailed in section 4, namely the co-identification of objects with a supervising agent in a crowded environment. The human supervisor hence identifies an object by looking at it, and the agent attempts to bring the object into the centre of its visual field on the basis of the supervisor's gaze direction and the discriminatory potentiality of the visual background. Nagai et al. bootstrap this process by embedded only two key mechanisms in the robot: visual attention and learning with self-evaluation. The former mechanism thus attends to objects that it considers salient within the robot's field of view, and the latter mechanism determines the success of the agent's goals on a statistical basis.

Nagai et al. are able to demonstrate that joint attention then emerges as a higher-level cognitive concept from these initial bootstrap mechanisms, and postulate that human infants learn joint attention (which is an essential prior stage in communication) in the same manner.

A more complex framework for autonomous learning in audiovisual environments, one that employs inductive logic programing to establish complex temporal protocols, is given in [118] and [76]. Here, symbolic descriptions of a continuous camera-and-microphone-based sensory feed are generated via unsupervised clustering. These symbolic descriptions are then passed to a logical induction system in order to generate protocol (rule-based) models. The system is hence capable of inferring the rules of simple visual games such as 'snap' and 'paper-scissors-stone', finally substituting a simulated composite of the supervising human agent into the visual setting to act out the inferred rules in a convincing fashion.

PROGOL is hence employed to provide an appropriately generalised first-order logical model of the underlying scene descriptors via inverse entailment. Should these symbolic descriptors be too numerous then redundant logical concepts with identical protocol relations will emerge, significantly increasing processing time in the inductive logic model, and requiring more training examples for accurate characterisation. Should too few classes be generated, however, the fine structure of the logical protocols necessary for complete characterisation of the game rules will not be available. Hence, some method of co-evolving logical inference and symbolic description is required.

What makes the system of [118] and [76] one of cognitive bootstrapping is thus the feedback that exists between the inductive logic module and the unsupervised clustering of feature attributes. By computing equivalence classes of objects related by identical logical structure, the inductive logical model provides a meta-clustering imperative within the attribute domain. The system is thus, to a degree, self-foundational in its percept categorisations; the initial bootstrap percept clusterings can be overridden in the light of new perceptual inferences. (There is not, however, as yet a feature disambiguation instruction arising from the inverse entailment; clustering is done with a sufficiently high bias towards over-fitting in order to guarantee all of the logical fine structure is represented at the outset). The system hence constitutes a two-stage realisation of the percept-action hierarchy implicit in self-grounding cognitive bootstrapping.

**Self-Modelling Approaches** Weng [144], in focusing on the problem of autonomous cognitive development, introduces a novel form of robotic agent, the *Self-Aware, Self-
**Effecting Agent.** As well as sensing and reacting to the environment via the usual percept-action structures, the robot modelling also includes percept states that relate to the internal structures of the agent, states, that is, that do not result in overt actions. The system thus has internal sensors and internal effectors, in addition to the usual external sensors and effectors.

**Awareness,** Weng proposes, is generated by an agent’s consideration of various action alternatives and their possible effects\(^\text{16}\). In order for an agent to improve its overall performance (in relation to some prespecified abstract criterion), it must therefore *model its own decision rules and their consequences* via the internal sensors and effectors. In this exercise in ‘learning how to learn’ the system hence becomes *self-aware.*

Weng goes on to apply this principle to cognitive vision (as realised in the mobile robot *SAIL*), envisioning a continuous sensory-motor architecture modelled on biological systems in which low-level pre-attentive vision is supervened upon by an attentive system generated by inhibitory signals from higher levels of the (layered) vision hierarchy. The system thus naturally allows for recognition of occluded images (for instance) by allowing selective focus on different parts of the simulated retina (which constitutes the lowest level of the vision hierarchy). The entire cognitive sensorimotor system is thus, in effect, a very generalised and hierarchical retina.

Hence, because there exists the capacity for the higher-level retinal components to inhibit the lower levels, and because the retinal hierarchy is pyramidal in structure, the upper levels effectively act as generalised supervisors of the lower levels. In this way, it is possible for *SAIL* to ‘chain’ learned responses together to form complex and novel behaviours from previously established lower-level responses. The system hence bootstraps high level cognitive categories from an initial pre-attentive world representation (which is simply the sensory input). The initial world representation thus has no goal-related aspects; these are dictated only at the higher-level, in response to environmental conditions (which can included supervisory aspects).

Weng’s model is hence very much in the connectionist paradigm of cognitive science; he argues that symbolic manipulation, if it arises at all, must occur within the sensorimotor system and that this alone is sufficient to account both for this as well as every other high-level reasoning task, such as planning and abstract reasoning. A very similar argument is put forward by Blank *et al.* in [9].

A further illustration of the importance of self-modelling in open-ended robotic agent learning is given by Buchsbaum [18], who proposes to utilise an agent’s own perception-action models to learn from another robotic agent’s experiences, without explicitly having to experience them for itself. She hence employs Simulation Theory, one of the dominant theories of cognitive agency in cognitive science, arguing that we make predictions about another agent’s mental states by extrapolating our own behavioural mechanisms into the other persons sensory space.

Simulation theory is corroborated by the biological discovery of ‘mirror neurons’ which similarly act as an interface between observed and agent-motivated behaviours (and are discussed in more detail in the final section).

In a similar approach, Polich and Gmytrasiewicz [107] use Kolmogorov complexity theory (discussed in the mathematics section) to allow an agent to model another agent’s

\(^\text{16}\) A view which may be compared to that of the phenomenologists in Part 1.
behaviour within a multi-agent environment via Bayesian updating. Using a Kolmogorov prior ensures that simplest models of other agents are favoured. Their method can thus, in principle, be extended to model the agent’s own behavioural responses in a compact fashion, giving rise to a limited self-model in the manner of Weng, allowing for the possibility of a fully autonomous cognitive agency able to assess and update its own learning capability.

7.1.2 Constrained Cognitive Bootstrapping in Autonomous Robots

Having looked at very open-ended forms of cognitive bootstrapping in autonomous agents, we now turn to more constrained forms. We use the word 'constrained' here to mean that the bootstrapping is limited in the cognitive categories the agent is finally expected to form; in particular being limited to graph and geometry-based models of its immediate environment.

Simultaneous Localisation and Mapping A particular sub-problem of robotic mapping that has received much attention is SLAM (simultaneous localisation and mapping), which relates to the potentially paradoxical situation of an agent attempting to determine its location within a map whilst simultaneously building that map (which in itself requires accurate location knowledge, hence the paradox). This is essentially a problem of cognitive bootstrapping, albeit within a very constrained sense: the agent does not have the freedom to open-endedly generate new cognitive categories; rather these are limited to particular types of geometrical or graph-based models of the environment, as well as positional and directional models of the agent itself. The principle cognitive category to be bootstrapped is thus the map-relative agent locality, which exhibits stochastic uncertainty in relation to both the ensemble of environmental map hypotheses, as well as over the individual map hypotheses themselves.

A detailed survey of the area is given by Thrun [135], who argues that the most successful approaches to this problem treat it in Bayesian terms (refer to the Mathematical Perspectives division for arguments as to why we might expect this to be so). Perhaps the most general mathematical form given for Bayesian mapping is that proposed by Endo and Arkin in [32]), which uses only sensorimotor information and proceeds as follows:

Given sensor outputs $s_i$ for a particular 'instance' $i$ initiated by the motor commands $m_i$, an 'event' may be described as the triple $e_i = \{s_i, m_i, n_i\}$, representing the particular moment that a robotic agent is considered (by whatever criterion) to have experienced an environmental novelty. $n_i$ is hence a tracking number, detailing the number of events to date (subject to possible memory limitations). The historical sequence of events leading up to and including $e_i$ is denoted by the superscripted quantity $e^t$ (and similarly for the quantities $m^t$ and $s^t$),

The posterior probability of the robotic agent re-experiencing the same event as that of the previous time $t'$ is hence given by:

$$p(e_t | s^t, m^t) = p(e_{t'} | s_t, s^{t-1}, m^t) \propto p(s_t, e_{t'}, s^{t-1}, m^t) p(e_{t'} | s^{t-1}, m^t)$$ (29)
\begin{align}
&\propto p(s_i | e_{\varphi}) p(e_{\varphi} | s_i^{-1}, m^i) \\
&\propto p(s_i | e_{\varphi}) \int_{e_{\varphi-1} \in \varphi} p(e_{\varphi} | s_i^{-1}, m^i, e_{\varphi-1}) p(e_{\varphi-1} | s_i^{-1}, m^i) de_{\varphi-1} \\
&\propto p(s_i | e_{\varphi}) \int_{e_{\varphi-1} \in \varphi} p(e_{\varphi} | m_i, e_{\varphi-1}) p(e_{\varphi-1} | s_i^{-1}, m^i) de_{\varphi-1} \\
&\propto p(s_i | e_{\varphi}) \exp(-RMS(s_i - s_{\varphi}))
\end{align}

(Applying the Bayes rule and Markov assumption as appropriate).

This quantity then represents the prospect of the robot localising itself in terms of the cognitive map. The term \( p(s_i | e_{\varphi}) \) is called the perceptual model and is estimated via the difference between the current and previous environments:

\[ p(s_i | e_{\varphi}) \propto \exp(-RMS(s_i - s_{\varphi})) \]  

The motion model is then the term \( p(e_{\varphi} | m_i, e_{\varphi-1}) \) and may be estimated (on the assumption of every event having been stored in memory) by:

\[
p(e_{\varphi} | m_i, e_{\varphi-1}) = \max(n^{n_{e_{\varphi}-n_{e_{\varphi-1}}} \exp(-RMS(m_i - m_{\varphi})), 1/i) \quad \text{if } n_{e_{\varphi}} - n_{e_{\varphi-1}} > 0
\]

\[
= \frac{1}{n_i}
\]

\[ (n \text{ and } \lambda \text{ are normalisation and weighting constants, respectively}) \]

If the current event can then be related to the previous event \( e_{\varphi} \) beyond a certain probabilistic threshold (that varies in inverse proportion to \( i \)), the system can localise itself at that event on a maximum likelihood basis.

This sensor-driven Bayesian model does not, in itself, specify any particular strategy for initiating motor commands \( m_i \), and can thus serve as the basis for imposing any appropriate model of intentional high-level motor imperatives, such as that of Bayesian exploration as set out in section 6.1.1. In this way Endo and Arkin’s model can be made into an explicitly cognitive bootstrapping one, without any a priori imposition beyond a distinction between the sensor-space and the motor-space (the motor-space linkages are not subject to uncertainty, unlike the sensor-space linkages; however, the sensor primitives themselves are free from ambiguity).

In general, however, SLAM proceeds via the building of explicitly spatial environmental models, rather than compiling sensory-linkages. A significant issue in automated robot mapping is thus the building of an initial model which can be refined by exploration of the environment (exploration of the environment being initially carried out in terms of this model). This, again, invokes the notion of cognitive bootstrapping, in that we are presented with the apparent paradox of simultaneously updating both the percept-space (the map-relative sensor readings) and the object-space (the sensor-relative map hypotheses). The situation is only resolved with a ‘bootstrap’ perceptual class that has the status of an a priori truth (in this case the agent-relative sensor-geometry readings). This bootstrap perceptual class then gives rise to an initial bootstrap map
hypothesis with sufficient accuracy to permit itself to employed in order to achieve further convergence.

Thus, for instance, in addressing this issue, Sim and Dudek [126] propose to bootstrap an initial spatial representation of an unknown environment by applying self-organised maps to a set of monocular images. This provides sufficient registration data to give rise to a plausible approximation of the agent's environment, from which a convergent model can later be achieved.

**The Spatial-Semantic Hierarchy** Expanding on this principle, Kuipers [67] proposes that in order to build an initial model when an agent with unknown sensor configuration is deposited in an unknown environment, we must employ a *spatial semantic hierarchy*. Thus, the first task of the agent is to identify distinctive states (dstates) within the sensorimotor interaction by which it can orient itself. It thus intentionally moves towards optically distinctive places, places in which there is no aliasing (ambiguity) in the scene description. Doing so, it may thus be able to ascend to the next stage of the hierarchy, and physically localise itself within a set of discrete map locations and possible trajectories. This hierarchical movement thus represents an abstraction away from the the continuous sensorimotor space to a discrete state-space in which possible *actions* are represented.

The next stage of the hierarchy introduces the ontology of places, paths and regions and their connectivity, order and containment relations. This is the *topological level*, and is thus a symbolic generalisation, in that it represents the potentially infinite series of paths between points A and B by a minimal descriptor. This is hence the level of the spatial semantic hierarchy at which most *planning* decisions will occur. An optional metrical level ascribes distances to these topological representations, providing a global map within a single frame of reference.

The map formed by the spatial semantic hierarchy is thus the product of a bootstrap process of increasing cognitive generalisation moving up the hierarchy.

A direct application of this abstract bootstrap methodology to the problem of robot place recognition is given in [68], where unsupervised learning of clusters of similar percepts is used to build topological and casual maps. These maps, critically, have a feedback path to the clustering so that, for instance, cluster disambiguation can take place. Low-level decisions can thus be reevaluated in terms of higher-level decisions and vice versa, thereby meeting some of the requirements of a true cognitive bootstrap mechanism, in which higher-level perceptual inferences capable of driving higher-level motor imperatives are possible.

### 7.2 Non-Physically Embodied Cognitive Bootstrapping Mechanisms In Computer Science

It is apparent from our survey of cognitive science that a number of non-embodied cognitive-bootstrapping mechanisms exist within computer science, in which the domain of activity is *textual* (for instance [96]). One particular current application of these methods is in the bootstrapping of web-based ontologies, in which the bootstrapping agents are 'web-crawling' robots that seek to extract the essential relations between
concepts via textual inference.

In the model of [75] a core domain ontology is pre-engineered, which allows the web-
agent to infer new relations within that domain. Some of the relations will have the
potential to extend the bootstrap ontology into non-domain-specific areas. These new
ontology elements can hence in turn allow new domains to be explored, and further
non-domain specific relations to be extracted. In terms of the percept-action model of
cognitive bootstrapping, the percepts in this scenario are the web-pages as interpreted
via the current ontology: the actions are then the movements between particular web-
pages (which can be carried-out in a domain-intentional fashion). High level inferences
can thus drive the low-level action processes, just as in physically embodied cognitive
bootstrapping.

Also within the textual sphere, but implementing the percept-action cycle more ex-
plicitly, Osawa [100] proposes that real-time representation of the cognitive inference
process will allow for the flexible and open-ended controlling of subsumption architec-
tures. In order to demonstrate this principle, he implements a dialogue-based natural
language interface based on a two-layer perception-action model. Within this model, inner utterances are triggered by sensory inputs (the percepts, consisting of typed user
inputs). The inner utterances constitute temporal sequences of patterns from previous
dialogue examples that the agent recalls and displays within the inner world. If these
inner utterances then prompt actions, in the form of recalled patterns of querying, these
are carried out by the interface as real, rather than virtual textual emanations.

In this way the system can bootstrap increasingly complex perceptual representations of
the sensory inputs (as measured by the extent of the inner utterances which they trigger),
with an empirical perceptual validation mechanism supplied by the query-responses of
the supervising agent.

Other areas in which bootstrap ontological inference techniques have been applied are
image ontology [65] and gene ontology [141]. Along with the textual domain, these rep-
 resent areas in which the embodiment of perception-action architectures (which we have
established as necessary if there is to be a meaningful criterion for empirical validation
of perceptual updating) can be commuted to the lesser condition of localisation. In
other words, the cognitive agent need not necessarily exist at a particular location (in a
physical or abstract space), it must only perceive and act at that location, leaving meta-
perceptual architecture to be embodied elsewhere, if necessary. (Hence a web-based
agent may be 'located' at a particular web-domain [ie in a virtual space], but embodied
physically at a distinct geographical location).

7.3 Summary of Division

We have surveyed both the open-ended and constrained variants of the cognitive boot-
strapping mechanism as it has been implemented within the field of robotics and applied
computer science. The constrained form, limited to fixed perceptual-category modelling,
has been particularly successful in the context of simultaneous mapping and localisation,
which may be considered to serve as a paradigm for the overall approach of cognitive
bootstrapping. Here, a partial environmental map can be used as an orientational per-
ceptual category in order to bootstrap further refinements of that same map by serving
as an approximate guide by which the robotic agent can manoeuvre into areas in which it
can validate alternative map models. A SLAM-based robot thus employs bootstrapping to overcome the paradoxical problem of requiring an accurate map in order to determine its location, and an accurate location in order to build an effective map.

We have also established that the same mechanism can exist in a non-embodied form, in particular as an autonomous ontological extraction mechanism.
Part VI

7.4 Summary of Survey

In the first division of this survey, concerning the various philosophical perspectives on the notion of cognitive bootstrapping, we sought to address a problem which goes to the heart of its necessity: the problem of how it is possible for cognition (particularly a cognition capable of modifying its own cognitive categories) to know, accurately, anything that exists outside of itself. The original response was that of Descartes; that one must be sceptical of any such external conjectures, the only certainties being the cognitive agent’s own existence as a thinking entity, and the hence the consequent certainty of the process of thought. This notion much later resurfaces in the classical symbolist approach to cognitive science and artificial cognition, in which the most abstract and formalised entities were deemed to be the most certain, and hence the task of artificial cognition was inherently a 'top-down' one of connecting logical entities to the raw material of sensory-data.

In the same vein, Hume argued that external objects consisting of unities of percepts do not exist for cognitive agents except as conjectures, since they cannot, by definition, be realised (as only individual percepts can).

Kant overturned these notions by demonstrating that the assumption of the existence of external objects is a prior condition of cognition. The formal conditions of cognitive experience thus demand, but do not constrain, a range of individual object possibilities. These concepts, which can exist for cognition only as unperceivable relations between percepts, are validated (or not, as the case may be) by direct cognitive experience of individual percepts in relation to their proposed linkages.

These linkages correlate directly with the motor-space possibilities of the cognitive agent, which must hence remain a priori and independent of any process of cognitive updating. An autonomous cognitive agent can thus doubt its higher-level perceptions (such as the nature of a distant building), and may even update its perceptual machinery in order to accommodate this uncertainty, but what it cannot subject to doubt is the action-based validation mechanism that may be applied to this perception (such as, in this example, the ability to walk around the building in question in order to resolve any visual ambiguity in terms of the lower-level visual percepts). It is thus implicit in the Kantian view that meaningful cognitive updating can only exist for an agent in terms of an a priori percept-action structure; without it, all perceptual hypotheses, all ways of interpreting the world, become equally valid.

This, we argued, provides the basis for the fundamental cognitive bootstrap that operates between the cognitive agent and the noumenal world that exists beyond it. The minimal initial perceptual hypothesis from which we bootstrap higher levels of representation capable of overcoming the Cartesian gulf between object and percept is thus the set of formal a priori requirements of Kantian cognition that unite low-level percepts together in synthetic object unities whose existence is necessary, but whose form is conjectural, thereby grounding the agent’s perceptual framework in something that exists outside of itself. It thus becomes possible for an agent to conceive higher-level perceptual hypotheses while retaining a framework for their validation.

Following this exposition of the underlying mechanics of perceptual updating, we then
went on to look at the hermeneutic circle, the bootstrap mechanism by which a cognitive agent gives meaning to the objective world as a particular and situated entity. The cognitive agent goes about this in such a way as to transcend the subjectivity inherent in the concept of meaning via the use of an iterative process involving the projection of percept hypotheses back into the environment for validation/falsification. We then investigated how the hermeneutic circle specifically relates to a perception-action-based cognitive structure via the account of the phenomenologists.

Following this, in the Cognitive Science Perspectives division of the survey, we sought to give an account of empirical evidence for the existence of cognitive bootstrapping, finding examples of it within both the symbolist and connectionist paradigms. We argued that perceptual updating is meaningful only as a means of compressing the totality of sensory inputs. In the computational linguistics section, we investigated how the embodiment of a perceptually self-updating agent naturally gives rise to a language semantics and syntax when a self-modelling and communicative capacity is implicitly incorporated into its design, emphasised most particularly by Luc Steels' "Talking heads" experiment.

In the Mathematical Perspectives aspect of the survey we looked at the underlying statistical theory behind cognitive bootstrapping in the form of Bayesian exploration, finding its most general expression and a priori justification in algorithmic information theory. We also speculated, after Chaitin, on the possible existence of empirical models of mathematical progression, in which axiom selection is itself approached via a cognitive bootstrap.

In the Computer Science Perspectives division of the survey we looked at actualised implementations of the cognitive bootstrap mechanism, finding it in a constrained form within robotic simultaneous location and mapping, as well as in more general and open-ended models of artificial cognition, such as Granlund's cognitive architecture, wherein the Kantian a priori synthetic concepts can be modelled in a very general way as invariant subspaces.

Having completed our survey of the four principle areas in which the cognitive bootstrap mechanism is found, we are hence now in a position to give a final, categorical definition of cognitive bootstrapping:

7.5 The Emergent Definition of Cognitive Bootstrapping

At the outset of this survey the term cognitive bootstrapping was intentionally defined in a relatively loose fashion. With the four subject areas of philosophy, cognitive science, mathematics and computer science now surveyed, we are able to provide a definitive and generalised description that remains applicable across all of the fields:

Cognitive Bootstrapping is the iterative mechanism by which cognition can become self-founding without falling into Quine's ontological relativism, in which any world interpretation is equally valid. It thus iterates between interpretation (in which percept categories are applied to the world) and exploration (in which sensory-data that has the potential to clarify the validity of the conjectured percepts is sought). Cognitive bootstrapping hence constitutes a physically embodied form of the hermeneutic circle.
Critically, since the exploratory phase is conducted in terms of the existing and potentially invalid percept categories, the initial 'bootstrap' hypothesis must have a degree of a priori validity in order to allow progressive convergence on an 'objective' model. Furthermore, there must exist an a priori criterion of percept-hypothesis validation/falsification implicit in the bootstrap hypothesis (such as haptic contact in the case of autonomous visual-haptic robotics). These a priori percept categories (often taking the form of contact-sensing and motor-space feedback within physically-embodied cognitive entities) are thus not admissible to the perceptual updating procedure, and represent the sole limitations on the extent to which cognition can become self-determining (we may hence legitimately doubt the applicability of our visual interpretation of an object but not the fact of our haptic contact with it, or the motor-impulse involved in reaching out to it).

Cognitive bootstrapping is, we reiterate, thus far more than unsupervised model regression. Unsupervised learning does not, in itself, reinterpret input data in terms of the hypotheses founded upon them. Rather, it forms hypotheses concerning the input data on the assumption of their independence to this process. In other words, they are not perceptual hypotheses, since they do not change the nature of the perceived sensory data.

Hence, unsupervised learning does not involve overcoming the paradox inherent in constructing a cognitive agent with unlimited capacity for forming novel percept categories with which to view the world, which must nonetheless be able to perceive whether these categories are representative of the world. Overcoming the paradox by bootstrapping requires that we have an initial set of low-level percept categories that we must assume are 'correct', and then progress from there to higher-level categories via percept-hypothesis formation and action-based testing. This initial category set, we argue, is the set of Kantian synthetic a priori cognitive categories which provide a framework in which Popperian falsification of percept category hypotheses can be adequately formulated. We further argue that this retention of the possibility of percept falsification is the means by which it becomes possible to distinguish the percept (and by extrapolation, the perceiving subject) from the perceived object. Without this mechanism a perceiving subject could not distinguish internal and external states with any epistemological certainty.

The question then arises as to what constitutes the minimal a priori category set required for cognitive bootstrapping in the artificial cognitive domain the a priori cognitive categories at the heart of the cognitive bootstrap need not be structurally identical with those of humans. For instance, in a cognitive architecture such as Granlund's [47], rather than an object category being imposed a priori, we have instead the broader-based a priori notion of invariant percept subspaces from which compact and invariant symbolic entities of increasing hierarchical complexity can be progressively defined, including the synthetic category of 'object'.

The context of symbol-hypothesis falsification in this architecture is then the percept-action link coupled with an exploratory imperative (even a simple 'random walk' imperative will suffice). Thus, the architecture presumes that the output of symbol manipulation must always result in an action, the effectiveness of which the agent must determine from within the percept space (which itself incorporates the higher level symbolic entities). Hence, an action imperative derived at the symbolic level (for instance,
the placing of one particular object on top of another) can only be evaluated as having been carried-out successfully by utilising both the higher-level symbolic categories (since the imperative was formulated in these terms) and the lowest-level object category (since this provides the primary link between the symbolic layer and the a priori sensory level of which it is an invariant subspace category). The symbol system is thus always semantically grounded; the system can spontaneously form and evaluate the suitability of invariant categories (which are always hypothesised), subject only to the constraint that it can not re-evaluate the validity of the a priori sensory level, or the invariant subspace categorisation mechanism itself.

There are thus other possible a priori constraints than those identified by Kant that may underly the cognitive bootstrap process in artificial cognitive mechanisms. Other issues arise in relation to the definition of the concept of cognitive bootstrapping given above, which we discuss under their appropriate headings in the following two sections:

8 Discussion of Philosophical Issues

8.1 The Necessity of Bootstrapped Cognition and Embodied Agency

The argument of the Philosophical Perspectives division of this survey with respect to the fundamental necessity of cognitive bootstrapping may be summarised as follows:

The a priori constraints on the relationship that exists between cognition and the cognized entity were first set out by Kant. Primary amongst these constraints is the notion of the transcendental unity of apperception, which gives cognition its unity despite changing perceptions, and which is hence logically prior to all cognitive experience. It then follows from the existence of this unity that there must be objects that exist outside of cognition, but to which cognition refers. Kant proves this via a temporal argument: we are conscious of our own existence (the transcendental unity of apperception) as occurring at a determined point in time. Temporal determinations are relative and presuppose the possibility of permanency existing for perception. The permanent cannot only exist within the cognitive agent, since the agent determines its existence in time relative to the permanent. This possibility of the permanent object cannot then be only a cognitive representation whose referentiality outside of perception can never by guaranteed; cognition must, in fact, have the capability to refer to external objects as a necessary precondition of its being. A similar argument in spatial terms implies that cognitive entities must exist as perspectives on transcendental objects.

Cognition thus arises as a synthetic unity out of the fundamental underlying 'noumenal' material of world, conditioned by it, representing it, but also imposing upon it the a priori conditions of cognition. The 'objective' world accessible to cognition is thus dictated neither entirely by the percept categories imposed upon it, nor entirely by the underlying nature of the noumenal world that exists outside of perception. The objective world accessible to cognition is thus, like cognition, of an emergent character; moreover, it is necessarily co-eval with cognition, and cognition is necessarily embodied within it, having a distinct location in both time and space. The fact that cognitive representation must exist at a particular time and place and yet refer to extended spatial and temporal
stretches implies that its representation is **compact**, a principle that is reflected in all of our examples of artificial cognition (for instance Granlund's invariant subspaces, or Marshall's meta-perceptual clusters).

A further consequence of this necessarily localised embodiment of cognition in the objective world is that its perceptions must be necessarily subject to **empirical validation**; a cognitive agent contained entirely locally cannot, other than in terms of the *a priori* cognitive categories, constrain in advance the nature of its perceptions at other times and places. Moreover, since a cognitive agent determines that a perception has one particular quality and not another, it is thus implicitly presupposed that it might have chosen otherwise. Cognition thus inherently posits the **falsifiability** of its (*non-a priori*) perceptions (a principle that also applies to the transcendental unity of apperception; a cognitive entity must always be presented with the possibility of its own ceasing in order to be present to itself; cognition must therefore always actively sustain itself as a unity).

We have thus implicitly differentiated a subjective and an objective domain; cognitive agents must, of necessity, be aware both of their own sensations, and also the possibility of their own sensations **failing to represent** the external world.

Hence, in order for a cognitive agent's perceptions to meaningfully correspond to the objective world it must be capable of updating its perceptual categorisations in the light of their inapplicability. It must hence enact a strategy of continuous cognitive bootstrapping in order to co-preserve the notion of an embodied, localised agent with an internalised perceptual domain, simultaneously with the notion of an accurately represented objective 'outside' that is ontologically differentiated from the agent's perception of it (a divide that is unbridgeable in dualistic Cartesian thought).

Hence, an agent that is aware in the Kantian sense, must also, of necessity, be an embodied agent with a bootstrap mechanism for cognitive updating. Any bootstrap process must begin with an initiatory set of assumptions (the bootstrap hypothesis) in order to proceed; we argue here that this minimal set of assumptions is precisely the set of Kantian *a priori* categories underlying cognition.

This is then a distinctly stronger motivation for cognitive bootstrapping than that given at the outset of the survey (namely, that building a cognitive mechanism that fully replicates human perception can never be achieved by an *intentional* act because of the limitations imposed by Russell's paradox; rather we require a self-tutoring, bootstrap mechanism in order to achieve human-like cognition). In combination, the two motivations would suggest that bootstrapping is an *essential* concern of artificial cognition.

An alternative, non-*a priori*, view of the necessity of bootstrapped cognition has also emerged during this survey in relation to Darwinian natural selection. It was argued in section 3.2.1 that environmental selection pressures on replicating agents in a rapidly changing environment (relative to the evolution rate) will always tend to favour cognitive architectures that generalise to the greatest extent given their initial configuration, and evolve via a bootstrapping process toward a local minimum in the disparity between internal world model and biological-survival-relative reality. Human societal (as opposed to morphological) evolution certainly meets this criterion, with survival demands on human communities typically changing on generational, rather than evolutionary, time scales. Here, the means of replication of human behaviour and understanding is not gene-based (which would respond only very slowly to environmental pressures) but rather
meme-based, that is to say, replicated via linguistic communication, and is hence capable of far more rapid evolution (see [26]).

Not necessitated \textit{a priori}, this linguistic bootstrap mechanism for the updating of cognitive categories is hence a specifically human one (given the very limited or non-existent language capabilities of animals). Cognitive modelling of this structure from within the Connectionist school would thus presumably centre on the differences in human cortical structure with respect to the standard mammalian brain.

Thus, a range of arguments for the inherent necessity of cognitive bootstrap mechanisms within general cognition have been brought to bear. We now wish to look at the implications that they have for the specific issue of artificial cognition.

### 8.2 Implications for Artificial Cognition

We have, in cognitive bootstrapping, described a methodology by which autonomous, self-updating cognition can attain a criterion for \textit{objective} validation of proposed perceptual updates. We would now like to discuss whether this objectivity extends to the \textit{cognition} so derived. A central question is therefore whether Kant’s transcendental unity of apperception, though by definition not \textit{experiencing} as a distinct perception to the cognitive agent that it \textit{defines} (being rather the basis of perceptual experience in general), can nonetheless be a tangible object of experience for another cognitive agent.

It would seem that the answer must be in the affirmative; we distinguish cognitive from non-cognitive entities on the basis of their \textit{transcendence of particular spatial relationships}. Thus, we do not perceive a particular cognitive agent at position \( x \) responding to a phenomenon at position \( y \) as being merely a causal link between spatially distinct entities; rather we perceive it as a tangible \textit{cognitive response} to the event at \( y \). This distinction is, of course, an \textit{a priori} synthetic construct in the same way that “\textit{A causes B}” is different from “\textit{B follows A}”: that is, in a manner that is not in any way empirically derivable. The cognitive unity of a third party is thus always established by the \textit{observing} subject, and not by any inherent characteristic of the observed party, other than in terms of what is it is capable of manifesting spatially as being \textit{indicative} of a cognitive unity.

Cognitive bootstrapping, as a mechanism for unifying sensory data into referential percepts, we have established qualifies as an \textit{a priori} necessity for cognition. This process of sensory generalisation can, moreover, be imitated by artificial mechanisms, as we have demonstrated throughout the survey. It would therefore seem possible that we can construct an artificial mechanism capable of \textit{manifesting} this generalising capability in the objective domain, such that an observing human agent (that is, one equipped with a transcendental unity of its apperceptions) is capable of \textit{recognising} the general unity underlying its actions and attributing the synthetic category of ‘cognitive’ to them.

The question then immediately arises as to whether this \textit{manufactured} third party is then truly \textit{cognitive}. In one sense the answer does not greatly matter; even if human cognitive agency finally proves to be unmanufacturable in principle, its inherent criteria for recognising cognition in a second party is built on its \textit{objective correlates}; there is no way to distinguish between the two possibilities empirically. In terms of Wittgenstein’s latter thought [150], this functional equivalence in fact constitutes an \textit{identity}; there are simply no grounds other than the metaphysical for differentiating the two possibilities;
it is a distinction without a difference. Hence, if an agent appears to confirm to the
category of ‘cognitive’ then, quite simply, it is. Our normal, contextualised usage of
language is, for Wittgenstein, the only basis for meaning.

Conversely, if we were to be consciously aware of the rules that governed a constructed
agent’s behaviour and further, conscious of our having implemented these rules in the
agent, Wittgenstein’s criteria of cognitive awareness would appear to suggest that the
agent could not then be cognitive in the same sense as ourselves; we would instead
perceive the constructed agent as constrained to follow its intrinsic rule-set, and thus in
the category of entities that are acted-upon (by ourselves), rather than acting (in the
same manner as ourselves). The agent would thus appear more like an automaton than
a cognitive subject by our innate criteria of cognition.

If, however, the agent has undergone a prolonged period of cognitive bootstrapping, such
that we can no longer ascertain the rule-set that dictates its spatial behaviour, then,
by Wittgenstein’s criteria, we are justified in attributing to it the notion of cognition.
The concept of ‘cognition’ is thus something that can only be bestowed on an agent by
another agent that is incapable of anticipating its perception-action rule-structure.

Consequently, by this definition, any of the open-ended models of embodied bootstrapping
detailed in this survey can be considered cognitive on the proviso that they have
achieved the requisite level of complexity. Cognition, is hence, from this perspective an
emergent property.

8.3 Self-Objectiﬁving Cognitive Models

Many of the models of cognitive bootstrapping we have explored have incorporated the
concept of self-cognition, the agent’s implicit acknowledging of its own cognitive agency.
This is central to Kant’s conception of the cognitive agent: it is also central to many
post-Kantian interpretations (for instance, the Phenomenologists, especially [85]). We
also saw in Part 3 that several linguistic models regard the T token as the precursor to
any sort of meaningful causal agency.

Kant hence argues that since a cognitive entity must both unify its perceptual possibles
and be present to itself in relation to this unity, it must exist as a particular, situated
perspective upon the object-concept that represents the unity of its perceptual possibles.
Hence, a cognitive agent is aware of itself positionally, but does not consider itself as
being simply a position, since it must also be aware of other possible positions with
respect to the synthetic object.

A further consequence of the Kantian view is that, when the cognitive agent attempts
to modify its set of perceptual descriptions of the objective world, it must, in order
to be capable of falsifying this notion (and thereby assess its objectivity), be able to
implicitly form a ‘third-person’ view of its high-level perception-action expectations in
order to compare this with actual perceptions. Hence, an agent with open-ended cog-
nitive capabilities must, at some level, be able to objectify its higher-level perceptions
in terms the lower-level ones and thus form a self-conception. At a still more general

\footnote{Such as when inferring novel high-level cognitive categories like ‘crowds’, and allowing this to actively respecify lower-level categories, such as, for instance, appending an additional attribute to the previous understanding of ‘person’ to suggest a tendency to form crowds, such that, henceforth, any perception of a person implicitly incorporates the additional datum ‘in crowd’ or ‘not in crowd’.

85}
level, a possibility employed by a number of the surveyed artificial cognitive bootstrap methods is to explicitly model the agent's own learning process with a view to refining it: essentially allowing the agent to learn how to learn.

At the less explicit level of cognitive self-representation, Heidegger, Merleau-Ponty and Sartre, for the philosophers, and Dewey, Ricoeur and Lakoff for the cognitive scientists, suggest that the form of our perceptions is dictated by our action possibilities. Self-models are thus implicit in our perceptual categories at the outset.

The extent to which this latter form of phenomenological self-modelling is capable of artificial modelling is open to question; however, the concept of affordances (Part 3) would appear to provide one possible avenue of exploration.

An implication of the more limited form of Kantian self-modelling cognitive agency is that action designations must be made in percept-relative terms:

8.3.1 Relative verses Absolute Agency

To give a specific example of this notion, first consider its converse: suppose that an agent without this capability embodies a functional mapping between the set of its possible perceptions, \( \{ P_1, P_2, \ldots P_N \} \), and the set of its possible actions (of magnitude \( O(|N^2|) \)) that are capable of linking these percepts together. An agent that employs percept-relative action designations, on the other hand, can potentially apply an arbitrarily smaller set of linking actions (of magnitude, say, \( O(2n < N^2) \)) to the same perceptual set with the same consequences. Thus, if \( A_{xy} \) is the action linking percepts \( P_x \) and \( P_y \), the total action set in the former case is some subset of:

\[
\{ A_{1,2}, A_{1,3}, A_{1,4}, \ldots A_{1,N}, A_{2,1}, A_{2,3}, A_{2,4}, \ldots A_{1,N} ; \ldots A_{N,N} \},
\]

whereas in the latter case the action set is a subset of:

\[
\{ A_{x,x-n}, A_{x,x-n+1}, \ldots A_{x,x+1}, A_{x,x+2}, \ldots A_{x,x+n} \};
\]

(for \( n < x < N - n \) and appropriate percept labelling)

Thus, in a sense, the percept-relative agent acts at, or via, \( x \). When the percept states are explicitly positions in a space (as, indeed, they are required to be by Kant) and not the structure-less qualia illustrated here, this analogy becomes exact.

Hence, percept-relative agency automatically permits low-level self-modelling in that the percept \( x \) can simply be defined as the agent's current state. How this might be accomplished in the case of non-percept-relative action agency is less clear; attempting to employ the current percept state as a description of the agent (ie as an active potential) fails by virtue of the fact that the agent's actions in relation to this percept are unique to the situation, and the percept is thus absolutely uninformative in relation to them (and the action states are by themselves unperceivable).

A further issue is that, in the case of the non-percept-relative agent, any novel perception-action linkage requires that the whole perception-action map be inferred anew. A percept-relative agency, on the other hand, can apply its set of action possibilities to any existing percept in order to generate a new perceptual experience because percept and action are largely decoupled. Hence, if this percept turns out to be of a completely
novel kind, the agent is able to accommodate this without having to update its action set, or having to completely update its perception-action mapping (the existing mapping still has validity over the majority of the perceptual range). Percept-relative agency can thus actively explore the perceptual terrain. The only limitation on the agent's actions is hence the algebraic closure of the action set, which may or may not be finite (thus, for instance, a hypothetical mobile robot might have an action set consisting of the two action elements: 'Turn 1 degree clockwise from current direction.' and 'Move forward 1 cm from current position.'), giving a finite perceptual range in the latter action space, but a potentially infinite one in the former).

The Kantian requirement of the cognitive agent's localisation in relation to synthetic object unities would hence appear to imply the existence of a set of percept-relative action possibilities in order that the agent's current action possibilities be represented, for perception, via a single percept. In maintaining a non-relative percept description alongside the relative agency, it hence also retains a model of the objective world alongside its implicit self-model. Cognitive updating with respect to the agent’s action potential can thus be distinguished from updating with respect to the environmental model.

9 Discussion of Cognitive Science Issues

9.1 Overcoming the Problem of Perceptual Reference in Open-ended Cognition

It is evident from our survey of the cognitive science aspects of cognitive bootstrapping that any embodied agent equipped with a high degree of cognitive autonomy must immediately address the problem of perpetual reference; how can it be sure that its percept categories and their updates are meaningful in relation the environment? We shall here assess how cognitive bootstrapping addresses this point in terms if the two distinct stages of the process that we have identified:

9.1.1 First Stage: Construction of the Bootstrap Percept Hypothesis

The construction of novel percept categories in an autonomous agent requires a mechanism for grounding these concepts such that they have actual, referential content. This, we have established is only achievable via cognitive bootstrapping within an explicitly Kantian framework. Hence, the initial 'bootstrap' cognitive hypothesis by which the autonomous agent first perceives the world in which it is embodied must have some element that is not simply hypothesised to be representative of the objective domain, but is rather guaranteed a priori to be be representative.

Examples of this in human cognition are the prior relationships that exists between our perception of spatiality and the space in which objects are embedded, and, in semantic bootstrapping, the prior referentiality of the core vocabulary from which further word-definitions are derived, such as those which make possible learning new vocabulary from a dictionary with meaningful referential content, as opposed to merely tautological or permutative interrelationship.

We hence agree with Millikan [87] that the a priori representativity of the congenital
set of human percepts is granted via natural selection (so that, for instance, if human beings’ innate perception of *ingestibility* in relation to external objects did not, to some degree, correlate with those objects in the environment that met with their nutritional requirements, then the species would simply not have proved biologically viable in the long term). Any artificial autonomous agent needs a similar minimal set of guaranteed referential percept categories, but, in the absence of a framework of natural selection, these would have to be imposed by their creators.

So, to give a vision-based example, consider the notion that the two separate images of binocular vision intrinsically refer to a *composite* object (as opposed to being merely correlated sensor streams). Rather than this synthetic construct being determined by the *survival value* implicit in it (say, for hunting purposes), this notion would instead have to be imposed by the artificial agent’s creators at the *outlet*, given that it cannot be inferred empirically (without invoking some equivalent *a priori* construct for unifying discrete sensory elements within the the composite sensory space).

Thus, the referentiality of perception must be ensured at the outlet prior to any cognitive updating. The question then arises of how, within the confines of these Kantian restrictions, is open-ended cognitive development to be made possible. We therefore now turn to the question of how the bootstrap hypothesis may be modified while retaining this referentiality.

### 9.1.2 Second Stage: Updating the Bootstrap Percept Hypothesis

The question of how perceptual updating is possible is equivalent to the question of how a cognitive agent that perceives only *what is expressible in its own percept categorisations* is nonetheless capable of forming a conception of the referential accuracy of percepts that are not themselves given *a priori*. The answer hinges on the fact that any novel perceptual hypotheses must themselves *composed* of the existing *a priori* perceptual categories, so that any perceptual update corresponds to a conceivable situation in the extra-perceptual world.

However, while sufficient to ensure the referentiality of the the proposed perceptual categorisation, this compositional strategy does not itself constitute a criterion of *validity* for the inferred percepts: so far no one inferred perceptual category is favoured over another. For this to be the case we need a perceptual optimisation strategy and a perception-action link.

We have seen that, in general, the perceptual optimisation strategy adopted by biological and artificial agents is one of perceptual *compression*; we wish to reduce the total sensory stream into a relatively few significant data. This, however, is still not sufficient, in itself, to determine the appropriateness of a perceptual update - after all, it is always possible to map every percept to a single datum, giving maximal compression at the expense of all environmental information. Thus, any novel perceptual inference must be allied with an action complex *within which this perceptual inference is sustained*. Hence, a novel inferred perceptual category such as 'is on top of' corresponds immediately to an action complex of the form 'put on top of' by which the validity of the perceptual inference may be tested.

We therefore test hypothesised percept-action *linkages* by comparing the current object hypothesis (the external world as viewed via the candidate percept, existing as a set
of proposed percept-action linkages) with the actual percept encountered on applying exploratory actions. The cognitive mechanism is thus, within the a priori constraints, capable of becoming self-founded.

Furthermore, because the agent can validate novel percept-linkage hypotheses in terms of explorative actions built upon existing percept-linkage hypotheses, it can hence improve upon partially successful or incomplete models in an active fashion. The partial successfulness of the perceptual model need only be such as to allow for broadly-accurate exploratory impulses (for instance, the partial map in a simultaneous location and mapping exercise in autonomous robotics that is sufficient to allow manoeuvring into poorly-mapped areas in order to update the map). It is hence in this manner that the process is a bootstrap; we utilise a percept mechanism of unknown value in order to interpret the external world in such a way that we can gain sufficient information in order to evaluate the worth of that perception mechanism. If it proves insufficient to the task of gathering enough evidence to validate itself, then it automatically fails that validation. If it proves robust at collecting information, but this information contradicts the percept hypothesis on a few occasions, the hypothesis may still be retained until sufficient data is collected to form a novel perceptual inference.

Autonomous cognition can therefore simultaneously modify both its world model and its perceptual machinery via an iterative process of bootstrap updating that, by virtue of its prior grounding, attains a plausible convergence with respect to both ontological spheres.

9.2 Symbol Grounding and Intentionality; Hierarchically

We thus have argued that the a priori nature of the core perceptual 'bootstrap' categories implies that any new percept categorisations must be made in terms of those bootstrap categories, but that these new percept categorisations can in turn be treated as the basis for further categorisations. The percept space of an agent is thus hierarchical and bottom-up in nature; novel percepts hence remain grounded in the manner of Harnad [50].

For an autonomous cognitive agent, validation of these novel percept-action structures involves, we have also argued, exploration of the percept domains in the hierarchical level immediately preceding it. Thus, to give a concrete example, in learning visually how to play chess through interaction with a playing (but not explicitly supervising) agent, it is necessary to first formulate the concepts of chess-piece and square before the concept of valid chess move can be meaningfully inferred via exploratory experimentation in terms of the former perceptual categories (and so on to still higher-level conceptualisations like strategic-position).

For a fully autonomous agent we must suppose that the exploration of the percept domains in the highest hierarchical level is of a random, or, at least stochastically driven, nature. Only in this way can groupings (clusters or correlations) of higher level concepts be found via unsupervised techniques; the agent has absolutely no basis for knowing in advance what higher-level inferences may be valid and hence must explore a representative sample of the entire space.

What is of particular interest in this scenario is that while randomised actions at the lowest level of the percept-action hierarchy appear to be just that, randomised actions
at higher levels of the percept-action hierarchy appear *intentional* with respect to the lower levels. Thus, while formation of a relatively low-level concept such as 'chess-board' might involve moving chess-pieces around at random until the full extent of possible movement is established, formation of the concept of 'check-mate' would necessitate random explorations of particular combinations of previously learned *valid* chess moves (the concept of 'valid chess move' having been established by unsupervised learning in relation to the actions of the opposing chess-player, who, we assume, simply reverses invalid moves).

Hence, the random exploration of the higher-level percept-action domain implies intentionality at the lower level; the *random* selection of the move 'queen's bishop takes king's rook' from amongst the ensemble of possible valid moves implies the *intentional* movement of the bishop to the square occupied by the rook. Thus, an autonomous agent with no overall goal other than generalised exploration can form an enormous range of intentional *sub-goals* by virtue of the hierarchicality implicit in bootstrapped cognitive structure.

This notion of hierarchically-grounded intentionality would then correlate with the existence of the 'sleep-paralysis' mechanism in mammals. According to the activation-synthesis theory [34], during rapid eye movement (REM) sleep, randomised neuronal stimulation is applied to the pons area of the brain as part of its memory consolidation activity. This randomised activity is interpreted at the perceptual level as *dreaming*. Dreaming is hence experienced as high-level visual and auditory stimuli of same sort that occurs in waking life, albeit with an appropriately randomised structure. However, this imagery is not merely abstracted symbolism, being rather *hierarchically grounded* in the percept-action complex of the organism. We thus have an innate tendency to *act out* our response to the dream-stimuli in an intentional and physical manner. It is therefore necessary for the brain stem to actively prevent this motor stimulation from making the final connection from the lowest-level of the grounded hierarchy to the muscles: a failure of this mechanism results in the phenomenon of *sleep-walking*.

A similar example of the hierarchical grounding of higher-level visual percepts in low-level percept-action mappings occurs in the mirror-neurons of the primate premotor cortex. It is found [39] that these particular neurons fire in response *both* to motor actions performed by the primate, as well as to those same motor actions performed by other primates in *the observing primate's visual field*. The high level visual percepts corresponding to the observed action must thus be hierarchically grounded in the appropriate lower-level action states.

10 Conclusions

10.1 Summary of Requirements for Cognitive Bootstrapping in Artificial Cognitive Models

We have thus argued on a number of philosophical and scientific grounds that any plausible cognitive model is necessarily a bootstrapped one, a categoric definition of which was given in section 7.5. We have also argued that cognitive bootstrapping is a practical necessity if we are to build an entity capable of passing as cognitive on our own
terms (the only possible criterion according to Wittgenstein), since Russell’s paradox would appear to render it a formal impossibility that we could knowingly formulate the underlying rules (if such exist) of our own cognition. Moreover, according to the Wittgensteinian view, any entity governed by formal rules of which we were explicitly aware would not strike us as being cognitive, but rather as an automaton governed by those rules. The remaining possibility is thus for these cognitive ‘rules’ to emerge by interaction between the artificial agent and the environment, commencing from an initial bootstrap set of cognitive rules (for instance the partial set of cognitive rules that it is possible, as humans, to finitely formulate). Critically, this bootstrap set must incorporate sufficient a priori structure as to allow for the possibility of perceptual hypothesis validation via action-based consistency checking.

Discounting the possibility (discussed in detail in the Philosophical Perspectives division of the survey) that artificial cognition is not achievable, we shall now look at the formal requirements imposed on artificial cognitive agents by the understanding of cognitive bootstrapping at which we have arrived (and which are exemplified in various ways by the artificial agents of Part 5).

### 10.1.1 A Priori Requirements

It is apparent from our analysis of Kant that a 'blank slate' approach to autonomous cognition is not possible; we cannot simply collate 'pure', perceptually unmediated, empirical data and hope to infer categories inhering within it. Certain minimal categorical assumptions must be inherently built into perception in order to define it as such, for example causality, topology and temporality.

In terms of autonomous robotics these restrictions mean that (while it is never meaningful to consider pure sensory data free of all categories) we should not, in general, treat raw sensory data with the most open-ended machine learning approaches in order to generate novel perceptual primitives at the lowest level of the cognitive architecture. Thus, we would never seek, for instance, to embed an autonomous robot’s stereoscopic camera-feed consisting of $2 \times N$ pixels and $T$ temporal samples within a $2 \times N \times T$ vector-space and attempt unsupervised statistical pattern-recognition in pursuit of complete cognitive open-endedness. Rather, we would utilise the prior knowledge of three-dimensionality and stereoscopicality in order, not merely to reduce the problem dimensionality, but also to provide a meaningful framework for the incoming sensory data, in which falsification of novel percept hypotheses is possible. (We can only validate a conjectured percept class with respect to some lower-level perceptual domain whose validity is assumed throughout; not to do so makes perceptual comparison meaningless [as when attempting to compare stereoscopic object models without having made a prior assumption of three-dimensionality]).

Hence, while we seek to make artificial cognitive bootstrapping as open-ended as possible, we should not seek to achieve this at the level of either the motor-space or of the sensory data. Rather, the open-endedness should refer to the breadth of possible meta-categorisation of the sensory data. Thus, in the manner of Marr’s [78] three-level description of cognition, we seek to build a perceptual hierarchy. While the cognitive category 'object' may possibly be required a priori\(^8\) we can hence seek to build an

---

\(^8\)Or perhaps some other suitable compressive relationship between percepts, such as Granlund's
open-ended hierarchy of meta-concepts built from such base categories (such as 'object adjacency' or 'object containability'). Very often, these categories will require a higher level logical structure (for instance, the move from first to second order logics). Thus, truly open-ended cognition would perhaps require an agent to bootstrap novel forms of logical inference, as well as bootstrapping new entities within the logic.

This is beyond the capability of the inductive logic mechanisms discussed in section 4.3.1, and perhaps requires algorithmic induction methods such as the Universal search of Levin (section 6.2.4). Alternatively, it may be possible to avoid formal mechanism entirely, and simply rely on connectionistic associations between simple learning units without any clear collective logical descriptivity (section 3).

10.1.2 A Posteriori Requirements

Beyond the hypothesis generating structure, we also require a mechanism to assess perceptual hypothesis performance in relation to exploratory actions undertaken with respect to it. We have here advocated Bayesian theory, not simply because of its statistical justification, but also because of its ontological neutrality and equivalence (demonstrated in section 6.2.2) to minimum description length theory, and hence to algorithmic information theory, within which Occam's Razor finds a formal justification.

However, this is by no means the only option; almost any inferential, or partial Bayesian mechanism may be implemented. We would argue, though, that such systems do not have the same level of a priori justification or ontological freedom, and have therefore the greater likelihood of failing in novel environments (avoidance of which is one of major motivating factors of cognitive bootstrapping).

10.2 Formal List of Requirements for Cognitive Bootstrapping

We are now in a position to tabulate precisely the component requirements of an autonomous cognitive bootstrap system as follows:

**Necessary Inclusions**

1. **A Bootstrap Assumption**: an a priori pre-cognitive structure capable of guaranteeing the validity (referentiality) of the bootstrap entities from which initial percept hypotheses are composed.

   (For instance a minimal set of given word meanings from which sentential querying can proceed in order to build up a language lexicon, or, more fundamentally still, the formal Kantian a priori requirements of cognition).

2. **A Generalisation Mechanism**: a compression mechanism for forming synthetic percept and object hypotheses. This must intrinsically favour the greatest compression consistent with the sensory facts (Occam's Razor); see Section 6 for an a priori argument as to the validity of this argument in MDL terms.

   (An example is Granlund's invariant subspace method [47]).

---

invariant subspaces.
3. A Hypothesis Testing Structure, allowing the action-based evaluation of percept hypotheses against empirically well established lower-level percepts. This allows for the hierarchical building of valid percept-action hypothesis.

(Thus, for instance, a visually-equipped cognitive agent might inductively derive the laws of chess in terms of its cognitive models of individual chess pieces, these cognitive models having achieved the status of valid percept hypotheses, though themselves built on a lower perception-action level).

4. Hierarchical Feedback between percept stages: lower-level percepts have the capability to be re-valdated in terms of higher-level percept-driven actions, as well as vice-versa (this does not extend to the a priori structure, which must remain inviolate throughout). Moreover, the agent can intentionally guide its lower-level percept training via the inferred higher-levels.

(Thus, carrying out a high-level agent action imperative such as ‘making coffee’ simultaneously supplies additional training data for the lower-level percept category ‘cup’. Hence the system achieves robustness through continuously reinforcing its perceptual self-training).

5. Embodiment: the agent must perceptually categorise the world in which it is active (which inherently limits its environmental modelling potential, since it cannot fully model its own agency [12]). The agent must hence compactly represent the world in a localised fashion. (It is also an a priori requirement of Kant’s that cognition exists locally within an external world).

(Embodiment hence serves to solve the frame problem by limiting the range of possible percepts to those relevant to the agent’s goals - for instance, survival in a Darwinian, naturally-selective environment).

6. An Exploratory Imperative. This predominantly takes place at the highest level of the perception-action hierarchy, and is randomised in the most unconstrained form of cognitive bootstrapping. Other variants are possible - for instance an imperative to explore regions of the percept-action space in which data is poor or absent, or exploration in terms of very general goals, such as the long-term survival of the agent.

This is hence the mechanism motivating the percept-action hierarchy-building, as once-novel percept concepts become increasingly well-established.

Corollary Inclusions

7. Self Modelling - This is implicit to an extent in cognition; an agent must at least model itself as a particular positional perspective on the world as an a priori condition of Kantian cognition. It must also implicitly distinguish between internal perception and the external world in positing the falsifiability of percept hypotheses in respect to actions, and thus carries an implicit model of itself as a perceiving agent.

This implicit model can be made more explicit if necessary, for instance by projecting simplified self-models into the agent’s world-model in order to refine its learning strategy (examples of which are given in Part 5).

8. A Self-Founding Percept Domain. This is a direct consequence of the fact that the (non a priori aspects of) existing percept hypotheses may be overridden in the light
of novel perceptual inferences, even when these inferences were made on the basis of experimentation carried out in terms of the former percepts. The only limits on this process are the a priori percept-action validation criteria.

References


