Epistemic Constraints on Autonomous Symbolic Representation in Natural and Artificial Agents

David Windridge and Josef Kittler

School of Electronics and Physical Sciences, University of Surrey, Guildford,
Surrey, GU2 7XH, United Kingdom. D.Windridge@surrey.ac.uk

1 Introduction

The aim of the following is to examine the epistemic[1] constraints on self-updating cognition applicable to both artificial and biological agents. In particular, we consider the problem of how the autonomous updating of an embodied agent's perceptual framework in response to the perceived requirements of the environment can occur in a logically-consistent fashion, such the ability to validate the agent's representation of the environment is maintained throughout.

Thus, a cognitive agent employing a representational framework, $\boldsymbol{R}$, must, upon examination of a set of observations $\{\boldsymbol{o}\}$ relating to the agent's environment, be capable of undergoing spontaneous transition to an updated representational framework $\boldsymbol{R'}$ in which the environment observations are transformed into an alternative set of observables $\{\boldsymbol{o'}\}$ (of possibility differing cardinality) that are deemed to be more 'representative' of the environment via some appropriate criterion of representativity. The question then immediately arises of what form this criterion should take, given that the only access that the agent has to the environment in order to determine the representa-

[1]

ep-i-ste-mic:
1. Of, relating to, or involving knowledge; cognitive.
(The American Heritage©Dictionary of the English Language, 4th Ed. Houghton Mifflin Company, 2004).
2. Of, or relating to, epistemology
(WordNet 1.7.1. Princeton University, 2001).
[From the Greek *epistēmē*, knowledge]

e-pis-te-mol-o-gy:
1. The branch of philosophy that studies the nature of knowledge,
its presuppositions and foundations, and its extent and validity.
(The American Heritage©Dictionary of the English Language, 4th Ed. Houghton Mifflin Company, 2004).

tivity of representations is via those very same representations. In additional to this foundational issue, a further difficulty attaches to the fact that individual representations of the environment are themselves necessarily conjectural, such that even *within* an appropriate representational framework, $\boldsymbol{R}$, there is a question as to which particular observation, $\boldsymbol{o}$, is most applicable to the current situation (i.e. classical perceptual uncertainty).

Systems for cognitive updating hence exhibit the potential for ambiguity between perceptual representation and perceived objects unless a means can be found to ensure that the two domains of inference can be empirically related while at the same time maintaining their epistemological distinction. (An autonomous cognitive agent must simultaneously employ some fixed perceptual reference in order to validate environment hypotheses, and a fixed environment representation to validate a particular perceptual framework). We shall hence argue that the notion of cognitive updating is ill-founded unless there exists a framework in which representational hypotheses can be empirically falsified via exploratory activity in the same way as the world representations described in terms of these hypotheses.

By virtue of having adapted to changing environments, sufficiently-evolved natural organisms (those that are complex enough to be considered cognitive) have an implicitly updated framework for environmental representation in which these difficulties are overcome. In such organisms, representational grounding is thus, to a large extent, ensured by 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. However, in so far as representations are *learnable*, biological organisms must employ an alternative mechanism for ensuring that the way in which the world is represented remains consistent with their survival imperatives. In doing so they must hence also address the problem of perceptual meaningfulness that lies at the heart of attempts to create cognitively autonomous artificial agents. We term the activity of mechanisms capable of achieving this *cognitive bootstrapping.* The concept of cognitive bootstrapping is thus analogous to (and indeed, to the extent that word-concepts are cognitive representations, *exemplified by*) the practice 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 falsifiable questions concerning meaning of new words, and thereby achieve the ability to expand our vocabulary indefinitely.

To this end, we survey a range of artificial cognitive mechanisms that attempt to address the issue of representational updating, concluding that only embodied perception-action learners capable of hierarchically-abstracting this relationship in such a way as to be manipulable in relational/symbolic terms are able to meet the indicated epistemic requirements. (Perception-action learning agents may be characterized as those for which 'action precedes perception'; that is, agents for which inferred higher-level percept states are considered meaningful only insofar as they relate to the agent's actions). Such artificial systems limit higher-level symbolic learning to that which is immediately relevant to the agent, defining the external world in terms of an increasingly complex set of motor capabilities, with the objects of the world consequently being represented in terms of their *affordances*. Translated into a multi-agent language-learning context, this means that agents engaged in evolving a collective communicative structure can hence only derive a meaningful syntax in relation to a semantics grounded in their respective active (and collective interactive) capabilities in the environment.

This embodied approach to autonomous cognition thus addresses a number of difficulties associated with classical artificial intelligence (in which intelligence is primarily regarded only in terms of the manipulation of symbols of fixed referential content), in particular those of symbol grounding and logical framing. Hence, in asserting that autonomous cognition is meaningful only with regard to embodied agents with limited action capabilities, the study of artificial cognitive systems is brought within the domain of evolutionary systems and adaptive robotics. As such, we believe these developments are of considerable potential interest to biological researchers.

The structure of the chapter will therefore be as follows. We commence, in section 2, with a discussion of the epistemological constraints on symbolic representations via an examination of the necessary *a priori* aspects of cognition that must be retained throughout any putative updating of the perceptual framework in order for such updates to be considered empirically meaningful. We consider how such *a priori* representations arise within naturally evolved systems. We then, in section 3, introduce the notion of embodiment within the context of artificial cognitive systems, and indicate, with examples, how this approach has the potential to address the symbol grounding and framing problems associated with classical artificial intelligence via notions such as affordance. In section 4, we address the nature of evolving representational structures in embodied *communicative* agents. We indicate, in section 5, how *perception-action learning* can be employed to hierarchically infer a grounded symbol set in order to create fully autonomous artificial cognitive systems capable of dynamically updating their perceptual framework in relation to the requirements of their perceived environment. We then, in section 6, discuss the issues raised in determining the epistemic constraints on the symbolic abstraction of perception-action architectures applicable to both natural and artificial agents. We conclude by giving a concise summary of the requirements of a cognitive system if it is to be capable of bootstrapping symbolic representations in such a way as to meet these constraints.

2 Open-Ended Symbolic Representation: A Philosophical Perspective

A Priori Constraints on Cognitive Representation

The argument of the survey revolves around a central paradox: how can a cognitive agent capable of changing its perceptual framework (that is, its way of seeing the world) ever *validate* one particular set of perceptual representations over another? The concept of validation would appear (at least in humans) to depend on the *perception* of the inadequacy of one perceived entity in relation to another: however it is not obvious that a perceptual framework could ever itself be an object for perception. The problem is certainly not soluble in terms of either the Cartesian or Classical Empiricist [21] schools of philosophy, since the first claims cognitive agents cannot absolutely validate the existence of anything beyond their own perception (itself built on a framework of pure reason), and the second does not recognize the possibility of the perceptual mediation of the objective world (objects present themselves as they are 'in themselves' directly to cognition).

Kant [22], however, provides an alternative conceptual framework, asserting that cognition, as a matter of *a priori* necessity, *refers* to entities existing beyond of an agent's sensory domain. Percepts hence serve to mediate between agent and object, being crucial to their distinction as ontologically 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, neither are they simply reducible to percepts. Instead, external object concepts are accessible to the cognition agent as *ordering concepts* imposed on intuitions (singular, low-level sensory percepts). Objects, as we understand them, are thus not themselves singular *percepts*: they are (in Kant's terminology) *synthetic unities.*

Thus, despite being necessary *a priori*, object concepts are of an inherently hypothetical nature, existing beyond the immediate certainty of the sensory impressions, serving instead as hypothesized *linkages* between those impressions. Immediate sensory impressions thus refer directly to the external world *a priori* in a way that can not be subject to empirical testing (being rather its *condition*). The conditions underlying perception are thus neither logically true nor false; rather they must simply be *assumed* to be true in order for cognition to take place at all. It is hence this *a priori* limitation on the possible updating of cognitive representation that will serve both to sets bounds upon, as well as to ensure the empirical grounding of, the concept of cognitive bootstrapping.

Implicit within this understanding of cognition is hence the idea that sensory intuitions can be linked together via *actions*, actions inherently having the capability to *test* object hypotheses, falsifying those that do not have the relationship between sensory impressions and actions implicit in the object hypothesis. Thus, we might need to *walk around* an object in order to establish whether the entity progressively revealed to the senses conforms to our conception of the object. Actions thus serve to test the *consistency* of an observed sequence of unfolding sensory impressions with respect to the underlying object hypothesis (which, at an appropriately generalized level, is itself necessary to give unity to the immediate sensory impressions). Object concepts thus implicitly serve as *singular* expressions of the functional mapping between individual sensory impressions and agent actions, which, (since they are not simply the equivalent of these functions) are inherently *compressive* in nature (thus, to give an idealized example, specifying an object within a view-independent 3-D coordinate-space is far more compact than setting out the exhaustive set of possible 2-D planar views on that object).

Cognitive Bootstrapping Within Kantian *A Priori* Constraints

It might first appear that the strong Kantian emphasis on *a priori*-limited sensory representation leaves little room for the perceptual updating required of cognitive bootstrapping. However, this is not the case; a significant question arises with regard to object concepts achieving a high-level of empirical confirmation. Since high-level object *hypotheses* link lower-level percepts together in a conjectural unity, these could, in principle (when sufficiently empirically confirmed) serve as the basis for further synthetically-unified object-concepts. In this sense, the original object-concept has become equivalent to a perception, albeit at a higher hierarchical level. Thus, we might, for instance, regard a very familiar object seen from only one perspective as a sensory-impression in its own right, and not in fact as an object-hypothesis that might be falsified by experience. This object might then form the basis of a new object hypothesis (for instance, by using it as a reference point for navigation), such that the new object hypothesis assumes the old one as a pre-assumed (though not fully *a priori*) basis for cognition.

It is hence the argument of this chapter that within the Kantian object-validation framework it is possible 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 approach built on the assumption of the *a priori* referentiality of the lowest level of the agent's perceptual hierarchy. Correspondingly, we require the *a priori* consistency and relevance of the lowest-level of the agent's *motor* space (so that, for instance, a cognitive entity cannot meaningfully query the topology of its motor-space independently of that of its perceptual-space).

The possibility of empirical validation of high-level synthetic percept/action unities thus rests on an *a priori* sensorimotor foundation. It is then these higher-level synthetic unities that enable perceptual hypothesis validation experiments of the following form: (where A_x are high-level actions and O_n high-level observational states, or perhaps stochastic distributions over states.)

If perceptual hypothesis H_1 is true then, for all definable perceptual transitions $O_m \rightarrow O_n$ such that $O_m \neq O_n; O_m, O_n \in \{O\}$, there exists a unique $A_x \in \{A\}$ such that $\{\cup A_x\} \Leftrightarrow \{A\}$

Which retains an empirical contrast with *object* hypothesis validation experiments of the form:

If object hypothesis H_1' is true then performing action A_x will result in observation O_m.

The former perceptual hypothesis validation experiment hence attempts to determine whether the proposed high-level perceptual framework represents the proposed high-level actions in the most efficient (i.e. least redundant manner) possible. In order that a space exists in which to perform this test of perceptual compression, there must be an underlying *a priori* space of actions and perceptions available to the agent which are not themselves subject to hypothetical uncertainty. 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 high-level choice of perceptual hypothesis, H_n, by which the world is to be interpreted, it is not free to choose an alternative set of action primitives, $\{A'\}$, or an alternative set of sensory primitives, $\{O'\}$, upon which the higher-level $\{A\}$ and $\{O\}$ are based (e.g. $\{A\}$ and $\{O\}$ might be legitimately defined in terms of arbitrary functions of n-ary action/perception concatenations: $\{A|A \to n^{A'}\}$ and $\{O|O \to n^{O'}\}$). $\{A'\}$ and $\{O'\}$ are hence the terms upon which the perceptual validation criterion is implicitly constructed (and without which perceptual reinterpretation is completely unconstrained)[2].

The relationship between $[\{A'\}, \{O'\}]$ and $[\{A\}, \{O\}]$ is clearly recursive. It is therefore the object of the following to propose that if an agent's *a priori* perception-action can be hierarchically extrapolated in this manner, it will be possible to arrive at at a sufficiently abstracted perception-action relation such that there exists a concept of the *symbolic representation* of the world. We hence now look at the the subject of *hermeneutics*, the branch of philosophy that deals with the interpretation of symbols, and, ultimately, the mechanism of cognitive *understanding.*

The Hermeneutic Circle and Cognitive Bootstrapping

Hermeneutics emerged initially as the branch of philosophy that deals with the interpretation of texts, only later acquiring its interpretation as the branch of philosophy that concerns the mechanism of human *understanding.* Central to the latter school of hermeneutics is Dilthey's notion of the 'hermeneutic circle' by which symbols can acquire an *objective* meaning. Thus, in order, say, to arrive at a dictionary of word meanings for a corpus of ancient texts, one simply proposes any *a priori*-plausible initial set of symbol meanings (for instance, a core set of words in an ancient text of known meaning with modern-day meanings attributed to the remainder), and then carries-out a reading

[2] Obvious candidates for the sets $\{A_n'\}$ and $\{O_n'\}$ in human cognition are, respectively, the motor complex and the space of visually and kinesthetically-determined body-relative positions of prioperception. Candidates for the inferred $\{A_n\}$ and $\{O_n\}$ might be e.g. the intentional act of cutting and the perceptual grouping of knife-like objects as constituting a distinct class of agent-utilisable entities.

of the entire corpus of work on this basis in order to arrive at an overall interpretation. This collective understanding is then utilized to *reinterpret* the component texts in the context of the whole. These reinterpreted component texts are again utilized to arrive at a new extrapolated interpretation of the corpus, and so on (errors in the initial set can also be corrected to a strictly limited extent in this manner).

Hence, generalizing this idea into the domain of active agents, the hermeneutic circle involves, firstly, an iterated movement from the outward manifestation of actions to an assumption about their inner, symbolically-determined motivation, and, secondly, proceeding from this assumption back again to a predictive conjecture about the outward manifestation of agent behavior, in process of circular empirical refinement. It is hence tacitly understood (though not explained) by Dilthey that this reiteration will achieve a degree of convergence on a final, stable set of symbol meanings (convergence to stability being the only possible criteria of finality). In this latter hermeneutic context, the attribution of *meaning* to symbolic terms is thus dependent upon the *embodiment* of the symbol-manipulating agent within the objective world; meaning cannot be conferred simply by the manipulation of symbolic entities (without descending into semantic tautology, such as when attempting to derive the meaning of *every* word in a language using only dictionary definitions).

For Heidegger [19] this tendency to regard cognitive meaning as being supervened upon by the action possibilities of an agent 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, the 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 physical being is employed in the perception of the pen, rendering the notion of an abstract mental plane of representations underlying our perception entirely redundant. This notion also extended to the derivation of the objects of knowledge (Zuhanden) from the praxical knowledge of action. Objective knowledge is now an abstraction from practical knowledge, and not its precursor. 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 cognition. Only when intended actions fail to account for our actual percepts, do we, '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 cognitive agents - we only perceive our own active potentialities in the world unless these fail to be realized as expected. There is hence in Heidegger's conception of objective knowledge, an implicit *ontological hermeneutic circle*.

However, notwithstanding these arguments, the subject of *artificial cognition* has traditionally been founded on Cartesian assumptions (namely, that cognition is essentially the rational manipulation of symbols that cannot have objective meaning). The difficulty associated with non-embodied symbol representation has consequently created considerable philosophical argument.

Hermeneutics and the Possibility of Artificial Cognition

One of the more persistent critics of the idea of artificial cognition is Dreyfus [9], 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. 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 categorizations of mind. Artificial cognition might thus exist, but not in any systematically pre-formalizable way.

Suber [47] 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 determining a concrete, existing model. A very large degree of semantic ambiguity would therefore appear to be associated with any finitely formalizable 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 [53] who argues that the fallacy of cognitive objectivism (the view that cognition can be tangibly formalized) is caused by overly formal logical structure of early attempts at simulated cognition (for instance his own SHRDLU algorithm, which is capable of passing the Turing test for intelligent behavior provided queries are restricted to the very limited but complete ontology of it's internally represented world). 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 Gödel proposition [15]). Instead the *embodied* agent can only allocate finite and partial resources to comprehending the world[3]. This

[3] Following historical difficulties associated with Russell's paradox (i.e. when the logical consequences of set-self membership are explicitly considered), we have become used to questioning the admissibility of a finite system such as the human

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 (he was later [52] to reject the possibility that any physically existing device for formal symbol manipulation can have 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).

We thus conclude that it is possible, in principle, for natural and artificial systems to overcome the paradoxes associated with open-ended representation in classical cognition and implement an embodied hermeneutic circle for attaching meaning to spontaneously-generated symbols. We turn now to the *science* of autonomous symbol generation in natural and artificial agents.

3 The Epistemology of Symbol Generation Within Embodied Agents

From the perspective of cognitive science it is possible to give a rather different argument for the form that cognitive bootstrapping must take from

brain encompassing a complete self-representation within itself (particularly a demonstrable self-representation). However, partial or temporally-retrograde self-models would appear to be permissible, so that it is possible for a human-being to use a linguistic token 'I' meaningfully and accurately, or, on a computational level, to build mobile robots capable of building accurate models of their *position* in space, if not of their full internal state-space. Complete and immediate self-models are ruled-out completely, though (see for instance [4] for a discussion of the limits to self-observation under finite, Markovian and infinite state-space assumptions, and [3] under quantum-physical assumptions). The use of partial self-models is thus, in essence, to adopt the hierarchical solution of Russell to his paradox via the *theory of types*, in which sentential reference can only be made to individual entities on the lowest level of the hierarchy, with sentential references to *sentences* about individuals being made only on the level immediately above this, and so on. Complete, but temporally retrograde self-models, on the other hand, occur at each iteration of the Universal Turing Machine that implicitly attempts to emulate itself within the Halting problem.

These issues give further impetus to the notion of artificial cognitive bootstrapping: since the underlying mechanism of human cognition can not be *knowingly* expressed in a finite and formally complete manner by any human being, it can not therefore be directly implemented using conventional methods of computational engineering. A human-equivalent artificial cognitive capability can thus only be achieved via an evolving, self-updating design approach. This is, in effect, to transpose the negative conclusion of the Hilbert programme (the attempt, in the 1920s to construct, in advance, a formal axiomatization of *all* mathematics) from a mathematic context to that of cognitive science, where the *laws of cognition* are hence the quantity that is incapable of a provably - i.e. *knowingly*- complete analytic formulation.

that given above, but to arrive at exactly the same conclusion. In this case, the argument is framed in terms of the problem of *grounding* symbols employed by autonomous artificial agents. An autonomous cognitive agent is, by definition, one capable of adapting to its in environment in behavioral 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 generalization, inference and decision making in order to modify the initial perceptual symbol set in the light of novel forms of sensory data (an also the mechanisms of differentiation and analysis to validate modifications).

Any representation that is capable of abstract generalization is implicitly governed by protocols such as those of predicate logic. As such, the generalized entities must observe strictly formalized laws of interrelationship, and consequently, in abstracting the symbol set away from the original set of innate percept-behavioral pairings, there is a danger of them becoming detached from any intrinsic meaning in relation to the agent's environment. A related difficulty, known as the *frame problem* [27], also arises in such generalized formal domains; it is by no means clear which particular set of logical consequences (given the infinite number of possibilities) that the generalized 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 applicable to symbols, and the 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 [39] to a Wittgensteinian [54] 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 enforced 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 [30]. 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 Sipper's An Introduction to Artificial Life [44] 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 ([25], [12], [16]), 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 [24]. It is consequently apparent that classical A.I. approaches to artificial cognition were of only limited success in that they attempted to build high-level environmental representations *prior* to considering agent actions within this model, rather than allowing this representation to evolve via hierarchical abstraction of the *a priori* percept-action relation [6]. Representative priorities were thus specified in advance by the system-builder and not by the agent, meaning that an *autonomous* agent would have 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.

The issue of representation is thus of the first importance to cognitive science. A central historical concern of the field has consequently been to determine whether mental acts can 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 they are to be interpreted at a higher level in terms of representations or schema. These two schools are respectively labeled 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.

A central problem for symbolic interpretations of cognitive psychology is thus 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 [33] were the first both to posit and to propose a solution to this problem from the perspective of cognitive psychology, centering 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 biological agents expressing Newell and Simon's ideas. This aspect of the symbolic account was further brought out by Pinker and Bloom in the context of language evolution [37], 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. Biologically-based accounts of symbolic causality thus agree that the *representativity* of mental symbols is characterized 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 biological 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 empirically checked for their referencing ability in terms of the agent's attempts to survive within the environment (equating to the hypothesis verification stage). The initial bootstrap symbol set is thus arrived at contingently, but the iterative convergence of the symbol reference system rapidly removes all traces of its random origin until an appropriate representation is arrived at (if only asymptotically).

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 necessary, if Newell and Simon's notion of physical symbol systems is to be extended to symbolic inference mechanisms 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 later discussion. We note for the present, however, that the innate, naturally-selected physical symbol set serves effectively as an initial perceptual meaning hypothesis for cognitive bootstrapping.

In contrast to the formal mechanics of the Symbolic approach, Connectionist accounts seek to comprehend agent meaning attribution 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. An example of this is Complementary Reinforcement Back-Propagation (CRBP) training within artificial neural networks [26], which is proposed as way of achieving self-volitional behavior in robots through neuronal constraints alone. Marshall *et al.* thus conjecture that self-directed learning behavior comes about as the result of competing tensions, such as that between the compulsion to model existing perceptual states effectively and the compulsion to seek out novel states. The 'homeostasis' thus achieved allows the network to bootstrap increasingly complex behavior patterns. CRBP directly models this behavior by, in addition to allowing back-propagation to reinforce internal goals in the conventional manner, also allowing the *complement* of the goal state to serve as negative behavior reinforcement during back-propagation. The tension between these contrary goal imperatives is hence directly modeled within the neural network structure, forcing the agent to test cognitive models by deliberately seeking areas in which they break down, and thus to refine them.

A key milestone of the Connectionist approach was thus the demonstration of the Boolean-logic completeness of such neuronal aggregates via the *multi-layer* perceptron (MLP) model. However, the MLP model lacks Turing-completeness due to the absence 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 [10] that the standard McCulloch-Pitts net augmented with expandable memory *is* Turing-complete and hence capable of arbitrary formal-language manipulation. The Symbolic and Connectionist approaches had, for the first time, thus achieved a demonstrable equivalence. Gärdenfors [12] later constructed a propositional language system based on the theory of functional dynamics applied to (purely abstract) information states. A neural network that undergoes learning generalization of the Hebbian kind in response to new information is thus shown to perform an *inductive inference* of the kind recognized in formal logic. Hence the symbolic/connectionist equivalence is not simply an *interpretation* of the the underlying neural connectionist model; it has actual referential capability.

At a more general remove, another approach to unifying the symbolic and connectionist accounts, 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 [55], sees the essence of cognitive agency within the world as being the ability to represent the varied mass of sensory information in a compact (and thus, generalized) form. Hence, grammatical rules may be regarded as compressed expressions of *language* possibility, 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 generalizers (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 change of surviving and reproducing than their less efficiently-compressing relatives. Progressive generations thus increasingly enhance the likelihood of agents with ever more economized 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 generalized *representative* capabilities (which is to say towards mechanisms of ever more efficient compression of *non-specific* data). This is hence a fully open-ended cognitive bootstrapping mechanism - the continuous need of the species to which the agent belongs to compress general, previously unexperienced sensory data amounts to a process of perceptual *hypothesis formation*, since the generalizability 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-founding in a process akin to the *hermeneutic circle*. We now look more closely at the specific form that the perception-action relation must take in embodied agents.

The Embodied Perception-Action Relation in Cognitive Biology

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 both to phenomenology (as indicated in the Philosophy section), and also to several long-standing schools of cognitive science. (Dewey had argued as early as 1896 [8] that perception, thought and action must be considered as part of the same stratum). A paradigmatic example of action-based perception in cognitive science is given in the study of environmental *affordance*, a term first coined by James Gibson [13], and specified in [28] 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, are thus objective in the sense of being invariant to arbitrary shifts in interpretation. However, a complementarity is implicated between perceiver and perceived: the criterion of accuracy for perceptual representation now depends on the agent's ability to represent *its own* active possibilities, i.e its self-model. Related schematizations of embodied cognition include Lakoff's [23] argument that *reason* is itself patterned by the spatial awareness of agency. Glenberg [14] similarly argues that conceptualization is constrained by the structure of the environment, our bodies, and our memory capacity. On the applied side of cognitive science are the searches for neural correlates of embodied cognition, for instance Berlucchi and Aglioti's [1] 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. This model provides a reference frame that further 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 first defined and, secondly, a constructive engagement between the world and agent's world-representation in order to refine this model. This work, and others like it, thus serve to validate Piaget's [35] notion that higher cognitive functions have their roots in lower-level biological mechanisms.

A similar idea is expressed by Millikan [30] with regard to language and intentionality, arguing that *function* can only be attributed to an entity within a biological context. She 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 communicating agents, since an unbounded notion such as the concept of 'addition' could never be proven to be the same for both agents. For example, one agent's rule of addition might 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 the addition concept. Millikan's resolution of the paradox is to propose that natural selection serves to remove the latter formulation of the addition rule on 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 typically experience in their biological lifetime), but uses more computation to do so. Hence aggregate natural selection will favor the smallest generalization consistent with the biologically necessary referents (thus providing a basis for Occam's Razor).

Millikan's work thus overcomes the classical *problem of reference*, where the relation between percept and object appears to be arbitrary (we might, for instance, ask why we regard the perceptual class *animal* as a singular entity, rather than as a collection of organic sub-objects or as a subpart of a species-collective). Millikan argues 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 animals 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.

4 Linguistic Signification and Embodied Agency

Perhaps the most obvious manifestation of the autonomous learning of symbolic representations occurs in human language. In attempting communication with another cognitive entity, agents must necessarily find a representation of the *commonalities* of their experience prior to allocating exchangeable linguistic tokens capable of standing in for these representations. That is, we must abstract from our immediate perceptions in order to find that aspect of them that is accessible to a real or putative second entity embodying a similar perceptive capability. 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 impressions. 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 describe their cognitive world at an appropriate level of detail. How is it then possible, in a communicative context, for cognitive entities to establish a common symbolic representation of the world that goes beyond what is necessitated *a priori*?

Implicit in this idea is the formulation of a symbolic representation of the agent itself. Rohrer [41], 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 *de-relativizing* of experience in order to establish a common frame of reference. Perry [34], Bermudez [2] and Metzinger [29] also agree that cognitive self-awareness (as manifested by the linguistic token 'I') *requires* all communicating parties to have internal representations of both the world and of the various inter-communicating agents; in no other circumstances can one explicitly attribute *perceptions* to *oneself*. (Viezzer argues in [49] that the symbol grounding problem can only really be solved by modeling both the agent's world [at the perception/action level] *and* the agent's modeling of the world in order to permitting *genuine* representational updating by the agent). Pinker [36] argues that language derives from an initial cognitive orientation attributable to an active agent (so that, for example, the fundamental noun/verb split mimics the perception/action division), which then develops along more complex lines via a semantic bootstrapping mechanism.

Spontaneous Language Formation in Embodied Agents

The study of spontaneous language formation in simulated agents gains its philosophical imperative in consequence of the *symbol grounding problem* first enunciated by Harnad [17]. Harnad's thesis demands a semantic interpretation of formal symbol systems that transcends the (merely syntactic) interrelationships available to the symbolic manipulation system in question. The problem Harnad identifies is analogous to the learning of non-native languages in humans; this is much more meaningful when attempted *in situ* amongst other speakers of the language than when learned from a dictionary. 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.* Steels gives perhaps the paradigmatic demonstration of semantic grounding in the formation of language in the 'Talking Heads Experiment' [45], the motivation for which 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 coeval in this scenario; symbolic syntax arises *at the same time* as semantics.

The talking heads experiment hence consists in a pair of robotic agents each equipped with a video camera and a set of predetermined low-level feature descriptors that can be arbitrarily mapped to internally-generated words. One agent is initially designated the 'speaker', and the other the 'hearer'. The agents occupy an environment in which planar objects of various colors 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 world description that would arise if the roles of speaker and hearer were fixed). Word definitions are thus characterized in terms of combinations of *a priori* feature descriptors of a visual nature; for instance, color, horizontal object positions, vertical object positions, etc. For example, 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 word-descriptors of the form: **A: vertical−position** > **0.5; B: vertical−position** < **0.5**. Or, equivalently, by descriptors of the form: **A: red; B: blue**

There is hence no unambiguously 'correct' object word-representation 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 an *expanded* experimental 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 *color* as a 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 (perhaps correlating with the neonatal synaesthesia hypothesis [18]). Equally, the hearer would need to evolve word descriptions that incorporated spatial considerations only in order to distinguish all three objects within the *extended* scenario. Steels' achievement is consequently in demonstrating that lexical convergence between speaker and hearer does indeed occur. 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 modeling lies at the heart of the symbol/referent relation. 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 and indicative capability similar (in *a priori* terms) to it own; self-modeling of perceptual agency is thus implicit in the experimental scenario. In philosophical terms, the talking heads experiment embodies the Wittgensteinian (cf [54]) view of communicative activity as a 'language game' in which agents invent words and meanings during their interactions, and opposes the Quinean [39] view that sees language as a series of inductive abstractions of perceptual correlations between word and object.

5 Approaches to the Spontaneous Generation of Symbolic Representations in Embodied Artificial Agents

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 is further to *construct* a sensory model of (presumably previously unexperienced) environments, we are then implicitly in the realm of artificial embodied cognition or *cognitive robotics*. 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 modeling 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 a significant 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 low-level vision tasks.

One particular area of investigation that implicates the notion of cognitive bootstrapping occurs at the interface of visual and haptic perception (e.g. [42], [43]). When a mammalian agent interacts 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 ambiguity present in visual data (particularly the ambiguities of binocular scene reconstruction). Moreover, it appears that the mammalian brain achieves this Bayes-optimally. 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 implicit in these visual-haptic models can by seen as the most tangible basis on which to implement an artificial cognitive bootstrap mechanism. Perceptions are hence seen as environmental *hypotheses* while actions are *hypothesis validation steps*. More specifically, *vision* is to be 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.

The degree to which artificial 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*. However, this spatial representation need not necessarily occur at the lowest level of the vision hierarchy, a point that will become apparent in the following discussion.

Hierarchical Percept-Action Approaches to Cognitive Robotics

Hierarchical approaches to autonomous robotics were first proposed by Brooks in [5], who employed the term *subsumption architecture*. The assumption of such architectures is that agent abilities are arranged in *levels*, with higher-level competences incorporating lower-level competences. For instance, the ability to plan a route presupposes the ability to avoid obstacles. Higher architectural layers hence control the behavior of the lower via the mechanism of inhibition, allowing the possibility of open-ended development of the cognitive agent's responses. Brooks notes that different forms of environment representation are appropriate to the differing levels, and that these levels can be extended indefinitely; however, the possibility of autonomously abstracting these higher hierarchical levels *along with* an appropriate environment representation is not directly considered. For this, we require an *abstractable percept-action* architecture.

Modayil [31] 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 lower levels of the OPAL (Object Perception and Action Learner) architecture. Bootstrap learning thus allows the system to move from egocentric (view-centered) and allocentric (object-centered) sub-symbolic descriptions to symbolic object-based description by ascending a four-fold hierarchy; Individuation, Tracking, Image Description and Categorization. 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 via the abstraction of a perception-action architecture.

Granlund [16] 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 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. 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 perception. Thus, the potentially exponential complexity of the percept domain is limited by considering only those percepts directly related to actions, which consequently occupy a far smaller state-space. (An idea of the information-theoretic disparity between these two different types of environmental modeling, the agent-specific and the agent-non-specific, is found in [32]). In the absence of explicit scene-representation 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 optimization procedures that allow compact representation, and implicitly, therefore, *generalization*. 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 representation 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 generalization. These higher level action impulses themselves generate further training data at the lower levels, allowing for robust and adaptive learning across the whole of the hierarchical structure so formed.

These compact representations within the hierarchical percept-action structure are *symbols*, corresponding, for instance, to the symbols employed in verbal communication. Such communication might 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 *always* grounded. 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 by virtue of the ’filtering-down’ effect wherein action impulses in the high-level abstracted cognitive categories result in progressively more *contextualized* low-level actions. Only at the highest goal-setting level is there thus a requirement for environment representations that are completely logically self-consistent (such as a coarse-grained reconstruction of the three-dimensional volume in which the agent acts): lower hierarchical levels need only be *para-consistent.*

Sun [48], 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 generalize. Stein [46] also argues that goal-based behavior in cognitive robots should be considered, not only at an abstract symbolic level, but also at the *lowest sensorimotor levels.* Hence, in projecting a goal, a robotic agent should utilize *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 currently hypothesized world-model. ’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 world-model. Stein’s MetaToto hence *self-trains* its higher-level cognitive abilities using only its internal representations. There is perhaps 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. Such 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 (treated more completely in the discussion and conclusions section).

A framework for autonomous perception-action learning that employs inductive logic programming to establish environment protocols and bootstrap appropriate high-level symbolic representations is given by the author in [51]. For a generic sensor-actuator coupling placed within a specific environment, only certain of the set of possible actions will serve to alter the percept space in a consistent fashion. Hence, after randomized exploration and induction of the rules governing this action legitimacy, the cognitive system sets out to eliminate redundant perceptual predicates in the inferred clauses in order to express a new, higher-level percept-action correspondence in which its actions are *always* successful. Such higher level perception-action representation is always of a more symbolic and abstracted nature than the generic sensor-actuator coupling, ultimately defining an open-ended series of logically-described environmental affordances of a form appropriate to verbal communication.

6 Discussion and Conclusions

We have looked at the problem of autonomous symbol generation in a range of natural and artificial agents, and have identified the mechanism of 'cognitive bootstrapping' as a means of accomplishing this in a maximally open-ended and epistemologically-consistent fashion. Cognitive Bootstrapping is hence the iterative mechanism by which cognition can become *self-founding* without falling into Quine's ontological relativism [40], in which *any* world representation can be considered valid. The mechanism 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 form of the *hermeneutic circle* within a perception-action learning context.

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 visual perception of an object but not the fact of our haptic contact with it, or the muscle-articulations involved in reaching out to it).

We thus overcome 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 *hierarchically* 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 *a priori* cognitive categories capable of providing a framework in which Popperian [38] falsification of percept category hypotheses can be adequately formulated. Without this mechanism a perceiving subject could not distinguish internal perceptual and external object 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 underlying the cognitive bootstrap need not be structurally identical with those of humans. For instance, in a cognitive architecture such as Granlund's [16], rather than an *object* category being imposed *a priori*, we have instead the broader-based *a priori* notion of *invari-*

ant 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 actual or potential *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 utilizing *both* the higher-level symbolic categories (since the imperative was formulated in these terms) *and* the lowest-level object representation (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 *hypothesized*), subject only to the constraint that it can not re-evaluate the validity of the *a priori* sensory level, or the invariant subspace categorization mechanism itself.

In terms of biological agents, *a priori* environment representation proceeds via Darwinian natural selection. However, environmental selection pressures on replicating agents in a rapidly changing environment (relative to the evolution rate) will always tend to favor cognitive architectures that generalize to the greatest extent given their initial *a priori* configuration. Such agents must hence evolve via a bootstrap process toward a minimization of the disparity between the biological agent's internal world representation and the species-based survival imperatives imposed by the environment. Human societal (as opposed to genetic) evolution 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 behavior 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 [7]). We hence agree with Millikan [30] that the *a priori* representativity of congenital human percepts is granted via natural selection (so that, for instance, if human beings' innate perception of *ingestability* did not, to some degree, correlate with those objects in the environment that met with their nutritional requirements, then the species would not have proved biologically viable in the long term). Any artificial autonomous agent would similarly require a 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 designers, perhaps motivated along biological lines.

Given that the referentiality of perception must be ensured at the outset, the question then arose of how, within the confines of these Kantian restrictions, open-ended cognitive development is actually to be accomplished. We have seen that, in general, the perceptual optimization strategy adopted by biological and artificial agents is one of perceptual *compression*; the idea being 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 proposed 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*. We thus *utilize* 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. Any new percept categorizations must hence be made *in terms of* well-established or perceptual bootstrap categories, such that these new percept categorizations can in turn be treated as the basis for further categorizations in a hierarchical fashion. We thus always maintain a 'fall-back' mechanism for empirical validation, irrespective of the perceptual framework adopted. A consequence of this is that an autonomous agent with no overall goal other than randomized 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 [20], during rapid eye movement (REM) sleep, *randomized* neuronal stimulation is applied to the pons area of the brain as part of its memory consolidation activity. This randomized activity is interpreted at the perceptual level as *dreaming*. Dreaming is hence experienced as high-level visual and auditory stimuli of the same sort that occur in waking life, albeit with an appropriately randomized narrative component. However, this imagery is not merely abstracted symbolism, being rather *hierarchically grounded* in the percept-action complex of the organism. Mammals thus have an innate tendency to *act out* responses 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 further example of the hierarchical grounding of higher-level visual percepts in low-lever percept-action mappings occurs in the mirror-neurons of the primate premotor cortex [11]: these 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 intentional lower-level action states.

Conclusion

In conclusion, it is apparent from our analysis of Kant that a 'blank slate' approach to cognitive updating is not feasible. Certain minimal categorical assumptions must inhere in perception in order to define it as such, in distinction to the perceived environment. In terms of cognitive robotics these restrictions mean that agents are not simply free to apply arbitrary generalization techniques to the reinterpretation of raw sensory data in order to bootstrap novel perceptual primitives. By the same token, biological agents (e.g. humans) capable of autonomous cognitive updating must employ a certain degree of naturally-selected representative capability in order to serve as a basis for further updating of their representational framework.

Cognitive agents must hence initially characterize their active environment according to pre-specified imperatives (species-survival in the case of biological agents, but potentially more general imperatives for artificial agents). However, we have demonstrated that the perception-action relation is capable of hierarchical abstraction to the symbolic level, with higher-level representations validated in terms of the high-level actions implicit in them. The only limit on the ability of agents employing this approach to bootstrap new perceptual categorizations is then the retention of the *a priori* structures required to give an empirical validation criterion for both the updated representational frameworks as well as the environmental representations themselves.

Acknowledgment: The work presented here was supported by the the EU, grant COSPAL (IST-2003-004176). However, it does not necessarily represent the opinion of the European Community, and the European Community is not responsible for any use which may be made of its contents.
This chapter represents the biology-related aspect of arguments first outlined by the author in the University of Surrey technical report 'Cognitive Bootstrapping: A Survey of Bootstrap Mechanisms for Emergent Cognition' [50].

References

1. F. Berlucchi and S. Aglioti. The body in the brain: neural bases of corporeal awareness. *Trends in Neuroscience*, 20(5):60–564, 1997.
2. Jose Luis Bermudez. Non-conceptual self-consciousness and cognitive science. *Synthese*, (129):129–149, October 2001.
3. T. Breuer. The impossibility of exact state self-measurements. *Philosophy of Science*, 62:197–214, 1995.
4. T. Breuer. Limits to self-observation. In G. Massey M. Carrier, L. Ruetsche, editor, *Science at Century's End: Philosophical Questions on the Progress and Limits of Science*. Pittsburgh & Konstanz, University of Pittsburgh Press & Universitätsverlag Konstanz, 2000.
5. R. A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 14(23), April 1986.
6. R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

7. R. Dawkins. *The Selfish Gene (2nd ed.).* OUP, December 1989.
8. J. Dewey. The reflex arc concept in psychology. *The Psychological Review*, (3):356–370, 1896.
9. Hubert Dreyfus. *What Computers Can't Do.* New York: Harper and Row, 1972.
10. S. Franklin and M. Garzon. Neural computability. In O. Omidvar, editor, *Progress in Neural Networks*, volume 1. Ablex, 1991.
11. V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(2), Dec 1998.
12. P. Gärdenfors. How logic emerges from the dynamics of information. *Logic and Information Flow*, pages 49–77, 1994.
13. J. J. Gibson. *The ecological approach to visual perception.* Houghton-Mifflin, Boston, 1979.
14. A.M. Glenberg. What memory is for. *Behavioral and Brain Sciences*, 20(1):1–55, 1997.
15. K. Gödel. Über formal unentscheidbare sätze der principia math. & verwandter systeme. *I. Monatshefte für Mathematik und Physik*, (38):173–198, 1931.
16. G. Granlund. Organization of architectures for cognitive vision systems. In *Proceedings of Workshop on Cognitive Vision*, Schloss Dagstuhl, Germany, 2003.
17. S. Harnad. The symbol grounding problem. *Physica D*, (42):335–346, 1990.
18. J. E. Harrison and S. Baron-Cohen. *Synaesthesia: Classic and Contemporary Readings.* Blackwell Publishers, 1996.
19. M. Heidegger. *Being and Time.* Blackwell, 1996.
20. J. Hobson. *The Dreaming Brain.* Basic Books, New York, 1988.
21. D. Hume. *An Enquiry concerning Human Understanding.* Oxford University Press, Oxford/New York, 1999.
22. Immanuel Kant. *Critique of Pure Reason.* Cambridge University Press, 1999.
23. G. Lakoff and M Johnson. *Philosophy in the Flesh : The Embodied Mind and Its Challenge to Western Thought.* Harper Collins Publishers, 1999.
24. D. Magee, C. J. Needham, P. Santos, A. G. Cohn, and D. C. Hogg. Autonomous learning for a cognitive agent using continuous models and inductive logic programming from audio-visual input. In *Proc. of the AAAI Workshop on Anchoring Symbols to Sensor Data*, 2004.
25. D. Marr. *Vision: A Computational Approach.* Freeman & Co., San Fr., 1982.
26. J. Marshall, D. Blank, and L. Meeden. An emergent framework for self-motivation in developmental robotics. In *Proc. of the Third International Conference on Development and Learning (ICDL '04)*, Salk Inst., 2004.
27. J. McCarthy and P.J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, (4):463–502, 1969.
28. J. McGrenere and W. Ho. Affordances: Clarifying and evolving a concept. In *Proceedings of Graphics Interface 2000*, pages 179–186, Montreal, Canada, 2000.
29. Thomas Metzinger. Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, (2):353–393, 2003.
30. R. G. Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism.* The MIT Press; Reprint edition, December 1987.
31. J. Modayil. Bootstrap learning a perceptually grounded object ontology. Retr. 9/5/2005 http://www.cs.utexas.edu/users/modayil/modayil-proposal.pdf.
32. C. L. Nehaniv, D. Polani, K. Dautenhahn, R. te Boekhorst, and L Canamero. Meaningful information, sensor evolution, and the temporal horizon of embodied organisms. In Bedau Standish, Abbass, editor, *Artificial Life VIII*, pages 345–349. MIT Press, 2002.

33. A. Newell and H. Simon. The theory of human problem solving; reprinted in collins & smith (eds.). In *Readings in Cognitive Science, section 1.3.*, 1976.
34. John Perry. Myself and i. In Marcelo Stamm, editor, *Philosophie in Sythetisher Absicht*, pages 83–103. Stuttgart:Klett-Cotta, 1997.
35. J. Piaget. *Genetic Epistemology.* Columbia University Press, New York, 1970.
36. S. Pinker. *The Language Instinct: The New Science of Language and Mind.* Penguin Books Ltd, 1995. ISBN: 0140175296.
37. S. Pinker and P. Bloom. Natural language and natural selection. *Behavioural and Brain Sciences*, 13(4):707–784, 1990.
38. K. Popper. *The Logic of Scientific Discovery. (translation of Logik der Forschung).* Hutchinson, London, 1959.
39. W. V. O. Quine. *Word and Object.* NY: John Wiley and Sons, MIT, 1960.
40. W. V. O. Quine. *Ontological Relativity.* Columbia, 1977.
41. Tim. Rohrer. Pragmatism, ideology and embodiment: William james and the philosophical foundations of cognitive linguistics. In Sandriklogou and Dirven, editors, *Language and Ideology: Cognitive Theoretical Approaches*, pages 49–82. Amsterdam: John Benjamins, 2001.
42. J Saunders and D. C. Knill. Visual feedback control of hand movements. *J. of Neuroscience*, 24(13):3223–3234, 2004.
43. E. J. Schlicht and P. R. Schrater. Bayesian model for reaching and grasping peripheral and occluded targets. *Journal of Vision*, 3(9):261, 2003.
44. M. Sipper. An introduction to artificial life. *Explorations in Artificial Life (special issue of AI Expert)*, pages 4–8, September 1995.
45. L. Steels. The origins of syntax in visually grounded robotic agents. In M. Pollack, editor, *Proceedings of the 10th IJCAI, Nagoya*, pages 1632–1641. AAAI Press, Menlo-Park Ca., 1997.
46. L. A. Stein. Imagination and situated cognition. Technical Report A.I. Memo No. 27, MIT AI Laboratory, 1991.
47. Peter Suber. Mind and baud rate. e-print of the Phil. Dept., Earlham College, Retr. 13/6/2005 http://www.earlham.edu/ ~peters/writing/baudrate.htm.
48. R. Sun. Desiderata for cog. architectures. *Philosophical Psychology*, 17(3), 2004.
49. Manuela Viezzer. *Dynamic Ontologies or How to Build Agents That Can Change Their Mind.* PhD thesis, University of Birmingham, UK, May 2001.
50. D Windridge. Cognitive bootstrapping: A survey of bootstrap mechanisms for emergent cognition. Technical Report VSSP-TR-2/2005, CVSSP, The University of Surrey, Guildford, Surrey, GU2 7XH, UK, 2005.
51. D. Windridge and J. Kittler. Open-Ended Inference of Relational Representations in the COSPAL Perception-Action Architecture. In *Proc. of International Conf. on Machine Vision Applications (ICVS 2007)*, Germany, March 2007.
52. T. Winograd and F. Flores. *Understanding Computers and Cognition.* Addison-Wesley, Reading, MA, 1986.
53. Terry Winograd. What does it mean to understand language? *Cognitive Science*, 4(3):209–242, 1980. Reprinted in D. Norman (ed.), Perspectives on Cognitive Science, Ablex and Erlbaum Associates, 1981, 231-264.
54. L. Wittgenstein. *Philosophical investigations : the German text with a revised English translation by Ludwig Wittgenstein.* Oxford : Blackwell, 2001.
55. J G Wolff. Cognitive development as optimisation. In L Bolc, editor, *Computational Models of Learning*, pages 161–205, Heidelberg, 1987. Springer-Verlag.