Concept-Centered Knowledge Representation: A 'Middle-Out' Approach Fusing the Symbolic-Subsymbolic Divide

Pei Wang[1] and Kristinn R. Thórisson[2,3]

[1] Temple University, Philadelphia, PA, USA
ORC-ID:0000-0002-1066-0454 pei.wang@temple.edu

[2] CADIA, Dept. Comp. Sci., Reykjavik U., Reykjavik, Iceland
ORC-ID:0000-0003-3842-0564 thorisson@ru.is

[3] Icelandic Institute for Intelligent Machines, Reykjavik, Iceland

Abstract. A long-standing paradigmatic debate in artificial intelligence is between the so-called *symbolic* and *connectionist* (or 'sub-symbolic') approaches to knowledge representation. Both approaches aim for uncovering the principles of how general, domain-independent knowledge is structured, generated, and handled by autonomous intelligent agents. Yet each seems to work only for certain kinds of information. The approaches spring from disjoint methodological beginnings; the former is inspired by human introspection ("top-down"), the latter by brain substrates ("bottom-up"). Neither approach has led to a unified theory of general intelligence. Few researchers are fluent in both methodologies, as using either approach calls for significant time and effort easily spanning decades. As a result, progress towards theories of general intelligence have been held hostage. We propose to brake this deadlock with a third approach: *Concept-Centered Knowledge Representation* (CCKR). Based around *situated dynamic knowledge graph generation and management*, CCKR captures latent features inherent in conceptual graphs that prior approaches do not address and adds capabilities that we argue are necessary for, and offer a path to, general machine intelligence. Here we explain CCKR and present arguments for its claims, resting in part on the results from two implemented experimental systems, the Non-Axiomatic Reasoning System (NARS) and the Autonomous Empirical Reasoning Architecture (AERA).

Keywords: knowledge representation · concept-centered representation · symbolic · sub-symbolic · connectionist · cognitive architecture · artificial intelligence · general intelligence · autonomy.

1 The Topic of Representation

Knowledge representation has been a central topic of artificial intelligence (AI) from its beginnings. The main issue at stake is the question of how a controller can control a body[4] in a complex environment, for various purposes, where only a fraction of the environment (and itself) – variables, outcomes, solutions, goals, problems, etc. – can be known (observed, isolated, summarized, measured, etc.) at any moment. No matter what type of tasks or information is involved, an intelligent agent's knowledge about the domain it inhabits, and the problems and goals it faces, need to be available to be processed and used to take action. This calls for a *memory* and a *format* – a ***representation*** – for storing those memories as knowledge, that is, manipulable and editable information structures, to use for guiding future behavior and learning.

Given an aim of matching human cognition in numerous environments, an agent with artificial general intelligence (AGI) must ultimately be able to deal with the physical world, with its infinite potential for variety and variation. In contrast to engineering methods that carefully design both the environment and the controller/agent, as a couple, to achieve predefined ends using predefined means, here we are concerned with agents that are guaranteed to *not know* everything they need to know to get things done, and must thus *figure out stuff* for themselves [83] and *learn on the job*, to get better at setting and achieving goals. Furthermore, situations where a learning agent can only interact with (measure and affect) a fraction of the state space it finds itself in at any point in time – and throughout its lifetime – and its memory and processing power is too limited to keep all potentially relevant details in mind, even when known, and situations where it is impossible for an AI system's designers to provide a complete list of what should and must be known before the agent leaves the lab, are not just inevitable, they are the norm.

This is what defines our scope here: Autonomy, generality, and cumulative learning [85], in worlds as complex as the one humans live in. Why we should pick such a lofty and seemingly "impossible" scope is due simply to the fact that the vast majority of interesting problems, environments, tasks, and situations that we can think of – and for which we might want a *very smart* machine – all come with essentially those exact constraints. For this challenge, over-simplifying the requirements is an easy mistake to make, with the obvious risk of the research missing its mark, or in the very least delaying progress substantially.

Complex dynamic (non-random) worlds present vast amounts of information, and the role of learning is to systematically keep track of useful regularities[5] in a compact yet flexible format. It is the *form, formation*, and *use of information structures* resulting from such learning that is our focus here, in particular, with respect to attainment of *increased generality and autonomy* of the cognitive control mechanism—i.e. improvements in adaptation through informed

[4] From a control perspective, a controller's "body" is by definition demarcated by its sensorimotor operations and the substrate in which control processes are implemented, which define what it can in principle measure, affect, and know. For all practical purposes, the term can thus be interpreted widely – as in 'a controller plus its controlled plant' – or narrowly, as in 'a biological agent in its natural habitat.'

[5] That is, any regularly recurring patterns that can be reliably measured in the world (by the agent) and trusted for getting things done (by the agent) can be considered *useful to the agent*.

learning (as opposed to random experimentation), over a diverse set of tasks and environments. To do so, the knowledge representation must, in our view, directly address autonomous (a) re-formulation (selective improvement, editing), (b) strategic partial re-use, through compositionality, (e) incremental buildup and maintenance (cumulative learning), and (d) compaction / compression (selectively non-lossy and lossy), including strategic deletion (informed, selective forgetting).

We present a theory-guided [105, 84] high-level *concept-centered knowledge representation* (CCKR) methodology for how the above requirements can be addressed. The approach aims to take even one step further, superseding prior attempts by offering a more unified, comprehensive, consolidating theory of representation for generally intelligent autonomous agents capable of *autonomous explainable cumulative learning.* The paper rests on a conceptual analysis, generalization, and discussion of representational approaches taken in two ongoing engineering projects with a combined six decades of research focused on autonomy and generality, the Autonomous Empirical Reasoning Architecture[6] (AERA;[7] [62, 84, 82, 90]) and the Non-Axiomatic Reasoning System (NARS[8] [93, 97, 103, 106, 94]).

CCKR can be seen simultaneously as a theory of concepts and a theory of how to implement such mechanisms in a machine [106]. The approach rejects the top-down methodology of the symbolic stance and the draconian bottom-up approach of neural-based approaches; CCKR can be thought of as a 'middle-out'[9] approach that starts from the ingredients that a thinking mind uses to create, manipulate and manage concepts, where the "upwards" direction from this middle ground links to higher-level phenomena such as plans, goals, and meaning, and the "downward" direction links to whichever substrate the mind is implemented—be it neurons, silicon, or something else entirely.

This is our first systematic consolidation and presentation of a CCKR approach, and the first exposition and analysis of how it challenges, and goes beyond, strictly symbolic and connectionist approaches to knowledge representation. While sharing some similarities with more common schools of thought, CCKR is in many important aspects fundamentally different from most well-known approaches to knowledge representation, as our demonstrations here of its principles show.

Several distinct meanings of the term *concept* can be identified in everyday language. A prominent one is as a synonym or placeholder for a "category" of everyday phenomena, e.g. "the concept of *a chair*," "the concept of *fast*." Another (and closely related) role is for referencing immaterial ideas (for instance, mathematical concepts like 'perfect circle' and 'infinity'). On an intuitive level, our use of the term can be said to be fairly compatible with many such uses, but since our aim is to provide a more rigorous meaning to it than a standard

[6] Also called the 'Autocatalytic Endogenous Reflective Architecture' [60].

[7] http://www.openaera.org – *accessed Jan. 7th, 2025.*

[8] http://www.opennars.org – *accessed Jan. 7th, 2025.*

[9] Thanks to Mike Judge and his co-authors of Silicon Valley for proposing this concept.

dictionary definition, some correspondences with the vernacular are inevitably weakened or abandoned.

1.1 Scope & Organization

After a short overview of selected methodological approaches to knowledge representation, we review and analyze the main features of the so-called 'symbolic' and 'connectionist' schools, and prior attempts to combine them. Then our *concept-centered approach to knowledge representation* (CCKR) is presented. To show the realizations and implications of CCKR, NARS [103] and AERA [62] are described, compared, and contrasted. Both systems are based on a CCKR approach and rest on common foundational assumptions about intelligence and cognition, while their differences expose some of the space in which arguments about further details, assumptions, and specifications of the CCKR framework must play out. To stay focused on arguments backed up by evidence, we limit any claims, descriptions, and examples of CCKR to the *existing implementations* of these two systems.

We also look at arguments for how CCKR subsumes prior approaches and other conceptualizations of knowledge representation, and compare our approach to existing ones, highlighting the advantages of CCKR in the development of generally intelligent systems. Lastly, we discuss the open issues. As our focus in this paper is squarely on representational topics, the discussion will center first and foremost on issues considered necessary for that, rather than on how the representation may be used in various cognitive functions of AGI systems.

2 Knowledge Representation: Early History

The numerous approaches to knowledge representation explored in artificial intelligence (AI) have been influenced by several other disciplines including computer science, psychology, mathematics and logic, linguistics, philosophy, and neuroscience; here we will limit our overview to the first three.

2.1 The Road to Representation

Work on knowledge representation grew out of early AI research and the study of data structures in computer science (c.f. [53, 67]). The subject of representation is familiar to any computer programmer, as computers cannot do any computing without representing information in some way. Going well beyond common structures such as sets, lists, trees, and graphs,[10] AI calls for representations that support learning and reasoning. Artificial general intelligence (AGI) makes additional requirements, including that the learning extends to a wide range of topics,

[10] While many information structures can be transformed into each other, here we are not concerned with structural and implementation details but rather operational properties relevant to the conceptual design of their use for understanding cognition.

tasks, situations, and environments, and that the learner be capable of imbuing this acquired knowledge with meaning (for itself, its tasks, its owner, and/or its environment), making it increasingly autonomous. The higher one climbs up the "ladder of intelligence," towards increased generality and autonomy, the stricter such requirements seem to get. While the full list of requirements for AGI is still being debated [88], it seems likely to require representation schemes that go far beyond the well-defined expressiveness, versatility, automated organization and processing efficiency of traditional programming languages, implying self-generated meaningfulness and flexible modularity in support of that self-organization [59, 86, 89].

Early work by the pioneers of the field focused intently on representation as a key aspect of higher-level cognition. However, progress in AI proceeded slowly in the first few decades, made only more obvious in light of the frequent overoptimistic predictions of imminent near-term progress. Results from research on humans in cognitive science seemed only to confuse matters, in part due to a misalignment of short- and intermediate-term objectives [108].

2.2 Behaviorist Approaches

In both psychology and AI, attempts were made to outlaw representation and focus instead on the world an intelligent agent inhabits [12, 79, 109]. To this end, Brooks proposed "intelligence without representation," suggesting that representations were, in fact, wholly unnecessary, as "[i]t turns out to be better to use the world as its own model" [12, p. 140]. In support of this extreme theoretical stance, he proposed a method of networked augmented finite state machines (AFSMs) for constructing robot control mechanisms for (semi-structured[11]) physical environments [13, 11]. However, instead of being representation-free, as the intent and claim of that research was, subsumption-based systems in fact *baked in* the representations directly from the very beginning, in the AFSMs network structures themselves: An AFSM-based controller necessarily includes hard-wired sensing, acting, and goal structures that rely on (the designer's) assumptions about the agent's environment, and the agent's future interactions with it. (Perhaps a more appropriate name for would have been 'Assumption Architecture'). While the subsumption approach does indeed assume, unlike behaviorism, the existence of internal goal structures, these generally cannot change automatically through learning or adaptation after the systems are deployed, due to features inherent in the methodology, and the approach thus not only fails to meet one of the key requirements for (general) intelligence, i.e. the ability to change one's mind – to re-evaluate goals, derive new goals, and to abandon them in light of new evidence [102, 83] – it also fails to achieve what it set out to do, namely, to do away with "internal representation."

Like the behaviorist psychology movement [79, 109], the subsumption architecture shifted the representational discussion in AI research to the environment,

[11] The methodology was thoroughly tested in household robots, most famously the vacuum cleaners of iRobot, a company Brooks co-founded.

in the hope of simplifying the research related to the control system at its center, attempting – but inevitably failing – to free its very subject matter from the theoretical basis on which it in fact depended. No behaviorist-themed approach has succeeded, or in fact will ever succeed, in ostracizing representations from an agent whose behavior can be considered in any way goal-oriented: The issue will end up being a blind spot in the methodology, or the representations given an implicit and entangled incarnation in the control system's very design. The cost is a theoretical handicap and cognitive inflexibility of the resulting systems.

2.3 'Subsymbolic' Approaches

Early interest in engineered systems inspired by natural neural networks dates as far back as the early cybernetics research of Wiener et al. (cf. [67]). Historical highlights of the development of ANNs (in a broad sense) include parallel distributed processing (PDP) "connectionist" models [74, 80, 71], Hierarchical Temporal Memory (HTM; [32]) and most recently, Deep Learning [44, 75, 30].[12]

Minsky's Stochastic Neural Analog Reinforcement Calculator (SNARC [52]), and Rosenblatt's Perceptron [72], were the first demonstrators of how to implement what came to be called 'artificial neural networks' (ANNs). Minsky, however, soon abandoned the approach entirely, along with another founding father of the field, John McCarthy [48], on the grounds that it wasn't sufficiently explicit, preferring strictly symbolic approaches (c.f. footnote 19 on page 21).

A variant of connectionist approaches is the so-called *dynamical hypothesis* of cognition [24, 8] (see further discussion on its roots on page 23). Influenced by research on chaotic systems, it borrows numerous abstract concepts from physics such as attractors, phase-space, and limit-sets. We consider its philosophical stance to be relevant here for at least two reasons. Firstly, on the positive side, more than the other schools of thought, it makes *time* a first-class citizen of cognition, something we consider a necessary (but not sufficient) requirement for research on general intelligence. Secondly, and on the negative side, by borrowing concepts invented to explain the behavior of low-level non-cognitive physical processes, and positioning these as part of a serious proto-theory of (very) high-level properties of intelligent systems (mental operations and cognitive control), its conceptual underpinnings sorely lack an intermediate level of explanation and operation. While being relegated to a descriptive, narrative or inspirational role at best, a significant level gap nevertheless exists between its theoretical concepts and the phenomena it purports to explain. As a result, dynamical approaches are insufficient for guiding the construction of complex cognitive systems in any specific way, and cannot provide a viable AGI research methodology. This shortcoming is shared by a number of other approaches, including behaviorism and neurology. In the case of both neuroscience and dynamical approaches, we

[12] These approaches are 'sub'-symbolic in that they generally focus on information at smaller scales than, or schemes that are orthogonal to, concepts (in the vernacular meaning); probably the best known of such approaches, 'connectionist' ones depend on a distributed (fragmented) representation.

estimate this spatiotemporal gap that they fail to address to span several orders of magnitude. It is in part this gap that CCKR fills.

2.4 The need for explicit representation

Like the founders of AI research, we consider representation to be central research topic for the field. For any research program addressing general intelligence directly, representations cannot be left out, and nowhere is the argument for this made more clearly than in Conant and Ashby's *Good Regulator Theorem*, where the role of models is explicated on solid mathematical grounds[13] [15]. Their work makes clear that it is not only desirable but in fact *impossible* to create a working controller without representation.

3 Representation: Symbolic vs. Connectionist

In this section we review the symbolic–connectionist debate from our perspective and set the context for how our approach is an alternative to the two.

3.1 Connectionist Representation

The connectionist approach comes from the idea that intelligence can be obtained by building a brain-like neural network – an idea dating back more than 100 years – and has been fueled by the recent success in applying artificial neural networks (ANNs) to a wide variety of information-centric tasks. Ever since the cyberneticists' work on (natural) neural networks [50], the model followed has been based on a large number of simple units connecting to each other via weights that are tuned (during "training") using special algorithms. Besides the obvious (albeit negligible) inspiration from neuroscience, this approach has been heavily influenced by mathematics research.

The typical form of connectionist AI is an ANN that is trained over numerous iterations on a given dataset, during which the weights of the connections are tuned. After an ANN is trained, the network's input/output layers forms a mapping for commonalities extracted from the training data, while the operation of the intermediate (hidden) layers is determined by the role needed to support this overall input-output mapping, i.e., the function represented by the network. Consequently, a vector that forms an input or output layer represents a sensation pattern (input) or decision (output), while a vector from a hidden layer of the ANN does not explicitly represent anything outside the network. Instead,

[13] Models come in infinite forms for an infinite variety of purposes. In a learning controller they are *empirical* when they are based on *empirical experience* (that is, in a way that can be argued to be based in some direct or indirecty way on measurement—i.e. 'grounded.'). In Conant and Ashby's theorem they could be formal ones [15], but need not necessarily be, and in fact cannot be when the axioms are not known (in which case they are empirical). Unless otherwise noted, it is these kinds of models we mean when we use this term.

it contributes to overall mapping of the model in a way that, for a particular ANN, is difficult to explain in succinct and simple terms. For this reason, ANNs excel as "pre-cooked" automatic classifiers; when posing as controllers, however, their behavior is difficult to ensure and predict, especially on data at the fringes of, and outside, their training data.

Though the various types of ANNs differ in many aspects, they share the use of what is called a *distributed* knowledge representation, which is fundamentally different from the *local* representation used in most symbolic AI systems. In distributed representation, "[e]ach entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities" [34, p. 77]. In contrast, in the symbolic tradition, each symbol can only have a single specific reference at any time.

To make use of an ANN in a practical setting, a fully trained ANN is typically placed in a larger architecture where it performs the classifying function it was trained for, serving as the representational component of a control system made up of other components. While most would say that the ANN "holds" particular representations that allow it to perform a complex function that requires "knowledge," this representation is unchanging and unchangeable in light of new evidence or information, until it is re-trained by its designers. Insofar as the "knowledge"[14] of the ANN references objects and phenomena outside itself, the output layer can be seen to implement a symbol system. However, this implies that the input-output mapping performed by an ANN – its 'knowledge-based' classification function – is grounded in the outside environment only at its input and output *ends* (in a fixed manner), as specified by the function. The ANN does not generate or obtain such knowledge itself, autonomously, and this information has thus no meaning to the ANN, only to its designers. Because the representation that the ANN holds is allonomically determined, i.e. decided, defined, and interpreted by the engineers during input-output training, it has been engineered to mirror an input-output mapping in their (human) knowledge, in light of *their* goals—not the ANN's. Such knowledge cannot be re-assessed on the job, so to speak, by the system itself—its reliable modification, no matter how small, requires a 'recall and full reset' of the system as a whole.

Generally speaking, connectionist approaches favor data quantity and representational uniformity over representational explicitness. The resulting knowledge lacks structural compositionality (modularity), inspectability, and explainability, to name three serious limitations in the pursuit of general intelligence. This, along with the features described in the foregoing paragraph, is the main reason why ANNs are always embedded in a larger control architecture: It does

[14] We consider 'knowledge' to be 'information that can be used to compose actions in pursuit of an explicit or implicit goal'. While an ANN certainly holds information that can be used to act upon, the ways in which an ANN can act on its knowledge is highly limited (note that this is different from the scope of *input* that an ANN can accept – which is limited in other ways, too complex and removed from the purposes of this paper to discuss here).

not "know what to do with its own knowledge" beyond the function for which it has been designed, as it lacks any facility to provide meaning to its internal information units independent of this function. Furthermore, intermediate nodes and node groups (e.g. layers) in an ANN may or may not map to anything in particular outside the training data and/or in other knowledge sets (such as its human engineers' knowledge). The intermediate outputs of sub-groups of nodes in the ANN are therefore meaningless (and largely unpredictable) from the human users' – as well as its own – perspectives, and one reason why the runtime operation of an ANN is difficult to understand.[15]

In summary, in consideration of the topic of this paper, the connectionist approach of knowledge representation is characterized by the following properties:

1. The system's knowledge is represented as a network of connected units which is often considered as approximations of neurons in a brain.
2. The system implements a function that maps the status of the input units (as a vector imposed by the environment) to that of the output units (also a vector as the system's response to the environment).
3. The input and output units get their meaning from their correspondence to (sensory, verbal, or other kind of) stimuli and responses, while the other (internal or hidden) units have no external correspondence and their roles are revealed by their contributions to the overall function.
4. Many human concepts may correspond approximately to status (activations) patterns of the network, which is *distributed* among the units and connections, rather than *local* to a single unit or connection.

Compared to the alternative approaches, the connectionist approach has nevertheless several major advantages:

- The function can be learned from training data in a connectionist network with universal approximation power, and does not require detailed human design.
- The learning process has a certain level of tolerance to uncertainty in the training data.
- The knowledge needed to build the function does not require a precise description.

Despite its notable achievements, the connectionist approach has met many fundamental objections over the decades, most of which are still unanswered by today's ANN methodologies and implementations. We can summarize the major criticisms on the representational aspect as such:

- The distributed representation in ANNs is very different from how human knowledge is expressed, which is one reason why ANNs have an

[15] The end-to-end usage of a neural network still has the symbol grounding problem at its input/output layers, except when the data directly come from sensors or go to actuators. A hidden layer has no such problem, not because it has grounded meaning, but because it has no identifiable *meaning* to talk about at all.

explainability issue and adversarial examples [57], as well as difficulty of receiving human knowledge [68].
- Since knowledge is coded as numbers with a single kind of relation ("strength"), inspection on particular knowledge is not supported [54], which also leads to problems in local revision and cumulative learning [87].
- Since knowledge eventually takes the form of an end-to-end mapping, there is no identifiable components that recursively form constructions. In general, it means that ANNs still lack compositionality, hierarchy and systematicity as criticized decades ago [82, 22].
- In input/output mappings, there is no distinction between correlation and causation [65, 91].

No doubt, the listed objections differs in severity and it may be argued that solutions to some of the above objections, which date back to 1986 or even earlier, are already at hand. Indeed, some *ad hoc* solutions exists for special cases of the above limitations. Every such solution, however, is only partial and comes with severe limitations. Most scathingly for research in general intelligence, there exists *no solution that addresses all of them together*, or even some notable proportion, and this is by far the most serious limitation of all such work to date.

3.2 Symbolic Representation

Historically, what has been called a "symbolic approach" to representation can be traced to a focus on human cognition. Meeting a need to constrain the discussion of human cognition in some way, a narrow focus on board games centered on the widely-articulated hypothesis that the game of chess could only be successfully performed by a general intelligence.[16] Looking at the history of AI, this hypothesis could not have been more wrong. The hypothesis and approach comes from the (arguably somewhat intuitively compelling) idea that AI needs human-like knowledge representation to solve complex problems. For the most part it meant that symbols were used to represent external objects and events, which were then to be manipulated via rule-based symbolic processes. The approach, also known as GOFAI ("good old-fashioned artificial intelligence") [31], rests on a tradition tracing back to the study of logic [40], psychology [66], linguistics [14], and philosophy [21].

In AI, one of the best known incarnation of this view is Newell and Simon's *physical symbol system hypothesis*: "A physical symbol system has the necessary

[16] The claim of Russian mathematician Alexander Kronrod in 1965, that chess 'is the drosophilia of AI' [20], may have sealed this particular board game as a major measuring stick on the advancement of AI, and quite possibly the focus on board games in general. Another reason for the field's obsession with chess was the hypothesis of Carnegie-Mellon researchers Newell and Simon [56] that a machine capable of winning a human world champion in chess would inevitably, by default, have obtained general intelligence, because that is required for playing chess—a hypothesis that could hardly have been more incorrect.

and sufficient means for general intelligent action." [56, p. 116].[17] It uses the concept of a "symbol" within a system to represent an "object" or "relation" outside the system, with the two connected by designation and interpretation. With such a representation, human knowledge in a domain, such as physics or biology, could be used by a computing system [33].

In this tradition, the symbolic representation was largely implemented in some sort of practical programmable logic, including first-order logic and extensions thereof, and resulted in a fundamental emphasis on formal logic and related automatic reasoning. In this camp was John McCarthy's theory of intelligence, according to which people used 'commonsense reasoning' to solve problems [48]. This theoretical foundation resulted in several research programs that attempted to mechanize expert knowledge, since symbolic knowledge encoded as logic statements, coupled with the proper (hand-coded) heuristics, should suffice to mechanize thought in a particular domain, allowing a machine to take the role of a domain expert through the systematic manipulation of symbols. The scope this approach was then extended to all human knowledge, leading among other things to the Cyc "common sense database" [45], which remains one of the largest research programs in AI to date. While the work by Lenat and Cycorp[18] on the Cyc system is unique in the history of AI, it had the support of Minsky, who, like McCarthy, was a strong proponent of a symbolic AI and dismissive of all connectionist approaches [54].[19]

For the current discussion, the basics of this approach can be summarized as the following:

- A 'symbol' is a token, sign or pattern with no particular or intrinsic meaning.
- A symbol is a 'stand-in' or "pointer" to something else.
- Anything can serve as a symbol of anything else.
- A system can use a symbol to represent an arbitrary outside object or event.
- The symbol and what it represents are related by an interpretation.
- The processing of a symbol by the system is fully determined by its shape, independent of its interpretation.

[17] The full paragraph in question states: "*A physical symbol system has the necessary and sufficient means for general intelligent action. By 'necessary' we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By 'sufficient' we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity".*

[18] http://www.cycorp.com – *accessed on Apr. 12th, 2023.*

[19] In his keynote address at the 2005 annual conference of AAAI, attended by one of us (KRTh), Minsky stated that "statistical approaches to AI offer nothing under the hood to work with," preventing, among other things, reflective cognition. This view is echoed in his 2006 book *The Emotion Machine* [54].

It is often neglected that this usage of the word “symbol” does *not* exclude some binary strings in a computer system: If a binary string represents nothing in particular, or something that is not an outside object of event, then it is not a “symbol” in the above sense. In other words, a symbol, according to this definition, must symbolize an *external* entity or phenomenon, and can be seen as the name or label of something that exist outside of the system holding the symbol. It is only in this sense that it could be argued that there exist “non-symbolic” AI systems – even if they use binary strings or variable names in their implementation that do not “represent” or “symbolize” any outside object or event.

A major attraction of the symbolic approach to AI researchers is its similarity to how human knowledge is often represented in everyday speeches and writings, as well as mathematics, etc., and its apparent closeness to the knowledge-handling traditions of related disciplines such as psychology, linguistics, and logic.

In view of the goal of AI to create a general learner, this approach to knowledge representation has the following features:

MODULARITY:

- Relatively fine-grained knowledge representation (consisting of small tokens).
- Any token can act as a symbol.
- Can be used to compose larger units according to a small finite set of compositional rules.
- Allows modular construction and de-construction of knowledge.

The above makes the knowledge representation explicit:

TRANSPARENCY:

- The representation format is relatively explainable and understandable (by a system’s human developers).
- The content is inspectable at a fine level of granularity.
- The meaning of the representation is systematic and compositional.

Although the symbolic approach dominated AI research for decades in various forms, it has been criticized from a variety of philosophical, psychological, and pragmatic angles, including that:

- The symbolic approach ignores the dependency of knowledge on an agent’s body and context [4, 42, 25]. A symbol is meaningless until it is grounded in something other than other meaningless symbols [76, 29].
- The approach is too rigid and brittle, and fails to capture the fluidity and creativity of the human mind [37, 26, 39]. Human knowledge in many domains cannot be fully formalized, abstracted, or symbolized [18].
- Its processing is limited to deduction on abstract knowledge, and is far from covering practical human knowledge [51, 10], especially procedural knowledge [110].

- To manually code all relevant human knowledge into AI systems is unrealistic and inefficient [82].
- The representation languages developed in the GOFAI tradition do not properly address how the knowledge could be assembled autonomously by AI systems [84].

3.3 Variants & Hybrids

Besides typical symbolic and connectionist approaches in knowledge representation, there exist approaches that are hybrids, as well as variants, of the two. One example that could be considered a combination of both is the above-mentioned *dynamical hypothesis* of cognition [24, 8]. Dynamic-system representation sees the system's state as a point in a multidimensional space, and the running process as a trajectory in that space. In this way, it is similar to a connectionist approach. On the other hand, when each dimension is considered as a specific measurement, it also shares properties with symbolic approaches.

Another approach that is between the two paradigms is the *Slipnet* of the *Copycat* system [36], a network of fluid concepts that are neither symbols denoting external objects nor models of neurons that work as functions.

Many attempts have been made to integrate symbolic and connectionist representations in one system [16, 9, 47]. One school combines the two approaches as cognitive modules in an integrated architecture, such as CLARION [81], ACT-R [1], SOAR [41], and Sigma [73]. Another school explores "how principles of symbolic computation can be implemented by connectionist mechanisms and how subsymbolic computation can be described and analysed in logical terms" [9, p. 3] using "neurosymbolic" systems [23, 17]. Concrete ideas for how to do this includes for instance using a neural network with dynamic external memory [27] and combining deep learning with Cyc [47].

These attempts at reconciliation invariably present some interesting ideas and demonstrate well-defined benefits over and above the approaches they combine, but they are, without exception, limited to sub-problems or special cases of general cognition. The bigger question of architectural unification and extension up to human-level cognition remains largely unaddressed.

4 CCKR: Concept-Centered Knowledge Representation

We now introduce an approach to knowledge representation we refer to as *concept-centered knowledge representation* (CCKR). While compatible with many high-level conceptualization of concepts in cognitive science (cf. [5]), our approach is based on concretely implementable ideas for how to build machines with a capacity for dynamic concept creation and use, cumulative learning, and empirical reasoning—in short, autonomous cognitive operation. To make the descriptions concrete, two AGI-aspiring systems we developed separately, NARS

(Non-Axiomatic Reasoning System) [93, 97, 103] and AERA (Autonomous Empirical Reasoning Architecture)[20] [82, 62, 61], are used as examples of how CCKR can be realized.

Intuitively speaking, at the highest level of abstraction, this approach takes the knowledge repository (memory) of a knowledge-based control system to be a conceptual (latent) network or graph, where nodes are concepts and edges are conceptual relations. To be more specific, the core ideas are:

1. Each *concept* is an identifiable (but flexible) information structure within a system that recaps (models) a portion of the system's *experience*.
2. The *general meaning* of a concept to the system is determined by its (experienced or imagined) relations to other concepts. Each time a concept is used, usually only a small part of its general meaning is involved, forming the concept's *immediate, spatio-temporally-bound meaning.*[a]
3. New concepts and conceptual relations are *constructed* by the system itself, to better model its experience.

[a] This refers to a concept's *foundational meaning* to the concept's autonomous owner; for a theoretical definition, see Thórisson & Talevi [92].

Further unique features define our proposal, of course, but these are the most important ones at the highest level. We can now proceed to explain each of the three ideas one by one, each in a subsection.

4.1 The Notion of 'Concept' in CCKR

Like all fundamental notions, "concept" has many usages and interpretations [43]. Our usage in CCKR is as follows:

Definition of 'Concept' in CCKR

We consider any data structure or information item a ***concept*** if – for any cognitive system capable of creating new knowledge (i.e., actionable information), that systematically relates to its goals, existing knowledge, and situation – the information item:

1. Is a unit that can be uniquely recognized/isolated (by a cognitive system's processes) and
2. one that both abstractly and concretely summaries certain segments of the system's experience, and
3. can be accessed and manipulated (by a cognitive system's processes) for particular purposes and operations (overt and covert goals), and

[20] Also called the 'Autocatalytic Endogenous Reflective Architecture' [60].

4. can (uniquely, fully, or partially) be related to other comparable or different units (by cognitive processes operating on them).

Data about the history of manipulation and use of the above information structures is embedded with it, so as to help consolidate, group, associate, and access, the information in the future. In this view, a concept can be seen as any information on the path to becoming, or already being, a relatively stable subset of a knowledge base that serves a practical purpose in the cognitive system's operations. Concepts are thus both macro- and microsymbolic[21] entities that not fixed, but rather, information sets that are coupled, to varying degrees of *strength*, between themselves and other such sets; a set whose elements are loosely-coupled holds a relatively vague concept, a tightly-coupled one is "crisp" or "clear." By being experience-centered and initiated, knowledge structures of this kind are not very much like traditional semantic networks in any way but rather an ever-changing network of dynamically-coupled augmented-symbolic information structures that are constantly being honed through experience of their use.

Concepts are thus chiseled by interaction with an environment and a social cohort, and thus take some time to form; fleeting and contrived concepts are possible – e.g. 'winged miniature basketballs' – as are strong, immutable ones—'mother,' 'cat,' and 'wind,' to take some examples from human experience that would be predicted by CCKR. Concepts exist because a cognitive mind has use for them; they become entrenched because they prove their value and get repeated use over long periods of time, which also enables them to have richer relations to other ones, both due to similarities and dissimilarities. Due to their compositionality (i.e. being hierarchical compositions of smaller concepts and knowledge elements), and the need for adaptation to a particular situation (which often involves a number of unique elements, e.g. a new goal that an agent has never pursued before), they are created or decomposed on-demand by appropriate cognitive operations in situ.

According to our usage, a concept cannot be a "nameless"[22] process or event that simply runs its course within the agent, whose side effects are untraceable after they happen (e.g. like an activation pattern in a neural network that is not associated with a name or an identifiable trace). Nevertheless, the concrete form

[21] In compositional representational approaches, we use the term 'microsymbolic' to refer to *items* that are *parts of* symbols in such systems, yet can also stand on their own and be accessible to the system through reflection; conversely, 'macrosymbolic' refers to concepts that rely on and/or encapsulate other concepts.

[22] It cannot be 'nameless' in the sense that it must be recognizable, manipulable, and operationable as a *unit.* here are undoubtedly many ways in which this requirement may be achieved, so this does not mean that we must resort to GOFAI or naïvely symbolic approaches, since in this discussion a 'name' is not a 'symbol,' but more of a 'pointer.' This will become clearer in the following sections; see also Section 6.2, page 41.

and content of a concept can still be different in different systems, and different at different times, under these requirements.

NARS uses a formal language, *Narsese*, to represent its knowledge. The basic unit of the language is *term*, and each term normally refers to a *concept*, which is a data structure in the system. In the simplest form, a Narsese term is a string of characters from a given alphabet. If the term appears in the system's input and output communication, it corresponds to a convention followed by the systems participating the communication, and the same term provides a way (at least partially and temporarily) to associate the concepts in different systems. Similarly, a term can be a word or phrase in a human language.

NARS also has concepts associated with sensorimotor devices and operations. Designed as a general-purpose system, NARS is not equipped with a specific set of sensors and actuators, but allows any hardware or software to be registered at a sensorimotor interface of the system. Each registered device, as well as each executable command of the device, will be taken as a concept, with a unique term as its identifier. In this way, NARS can issue commands to a device to carry out an action and get results and feedback.

Narsese uses a set of term constructors that each can build compound terms of a certain type from the existing terms, recursively. A compound term, including a compound perceptual terms and an operational terms, also refers to a concept.

AERA takes a somewhat different approach to CCKR than NARS. Among its fundamental atomic information structures relevant here are *variables*, *composite states*, and *models* (composite and atomic).[23] Replicode, the language an AREA agent's knowledge is represented in [59], is late-binding throughout, and variables work both as placeholders, pointers, values, lists, and arrays; composite states are variable combinations, and causal-relational models (CRMs) are a kind of bi-directional copula that may represent any transformation or relation, an important type of which is, as the name implies, causal [91, 62]. CRMs have a left-hand side (LHS) and a right-hand side (RHS); to represent a transformation, the LHS is a prior state (typically represented with a composite state) and the RHS is the result of a transformation. The model holds the function for the transformation it describes (how the values on the LHS change to those on the RHS), along with the context that the model is relevant for.

A single CRM can itself be considered as a concept according to our definition, albeit a peewee one, from which other concepts are constructed. Most concepts in AERA are composed of a large number of models, and the complete set of such models is not fixed, as they are in part contextually defined. Their operational semantics emerge during the system's experience with using them in context, which ensures their grounding and forms the basis for the system's learning. For instance, the concept of a deictic gesture (pointing at something, e.g. with your finger) would – depending on an agent's experience – include sev-

[23] We restrict our discussion here to information structures that an AERA agent uses for autonomous learning; other information structures can be used as part of a programmer-defined *seed*, but this is a more involved subject that need not be address here.

eral models describing the shape of the hand, the relation between the object pointed to and the finger/hand, as well as how to time the occurrence of such a gesture with respect to speech (see e.g. [90]). The concept would not necessarily have any name associated with it, but rather would be summoned in part or in whole, as needed, through relevance mechanisms, depending on its intended usage in the current context (e.g. whether for generating a pointing gesture or for recognizing it when performed by someone else).

The knowledge that an experienced AERA agent has learned about a complex domain from the ground up will consist of a very large set of such peewee models [58]—the size of the set being a function of how knowledgeable the agent is.[24] The largest such knowledge base created from scratch through an agent's experience, as demonstrated to date, was our S1 agent that learned to conduct a spoken-language multimodal interview with a human by watching humans do it. This agent's knowledge started with 26 hand-coded bootstrapping models; after learning how to perform an interview it contained 1400 models [90], which it had created autonomously over a course of 20 hours of observing two people in realtime dialog.

CRMs can effectively encode causal relations ("affordances") such as "sittable-on," which allows them to be used effectively for plans and action. The concept of 'chair,' for instance, could have dozens, hundreds, or even thousands of AERA models—many of which would be shared with other concepts (CRMs natively support transfer learning mechanisms in AERA), but some of which would be unique to chairs. The couple of "sittable-on" and "single-person," each of which could be represented by a single CRM (but will then reference other models as well, mostly lower-level), will *together* be rather unique to chairs, even though sofas, benches and the side-walk can also be sat on (and thus considered 'seats' but not 'chairs'[25]).

4.2 The Meaning of a Concept in CCKR

According to CCKR, the bulk of a learning agent's memory, at a particular point in time, consists of a (implicit or explicit) concept network that as a whole contains the systems' total knowledge, and which, given the set of possible operations on them, defines the limits of everything that the systems can be said

[24] Our knowledge representation method allows re-use of knowledge, considerably reducing representational needs for domains rich in self-similarity (in the limit, memory increase for learning an additional domain $\mathcal{D}_a$ at a single level of detail, $f_{mem\uparrow}(\mathcal{D}_a)$, is close to the memory requirements for additionally storing the difference between the new domain and the most similar one already learned, at that level of detail, $\mathcal{D}$, assuming perception/acquisition of their differences, $\Delta = \mathcal{P}(\mathcal{D} \setminus \mathcal{D}_a)$, approaches perfection: $\lim_{\Delta\to\infty} f_{mem\uparrow}(\mathcal{D}_a) = \mathcal{D}_a - \mathcal{D} + \alpha$, where α is a small fractional overhead).

[25] It should be noted that here we are using these words for human concepts purely for convenience's sake; an AERA agent that learns concepts from scratch will be entirely limited to its own world of experience and may thus not end up holding such human-centric concepts at all.

to "know" at the moment.[26] In this network, the nodes are concepts, and the links are conceptual relations. There is a number of basic link-types, each operating in accordance with particular semantically closed operations [64], that is, what such a link means is defined completely by what the system can do with the link, in accordance with our definition above (see p. 24). The other (non-basic) relations between concepts are concepts themselves, and they are related to the basic types.[27] The network changes constantly (in terms of its topological structure as well as the attributes of the nodes and links) when the system is running.

Given a conceptual network of this kind, the meaning of a concept to the system is determined by its relations with other concepts, and these relations have (almost exclusively) come from the system's *experience*, which consists of spatially localized streams of input and output over time. Each such stream consists of conceptual relations itself, including temporal and spatial relations as special cases. This way, the meaning of a concept in CCKR is "experience-grounded," and is accessible to the system, rather than given by the system's designer through pre-conceived notions, or by an interpretation that maps it to outside entities.

In NARS, the basic conceptual relations are the *copulas* that intuitively link two terms into a statement. There are four of them:

Inheritance: $\mathsf{S} \rightarrow \mathsf{P}$ means that "S is a type of P."
Similarity: $\mathsf{S} \leftrightarrow \mathsf{P}$ means that "S is similar to P."
Implication: $\mathsf{C} \Rightarrow \mathsf{R}$ means that "C implies R."
Equivalence: $\mathsf{C} \Leftrightarrow \mathsf{R}$ means that "C is equivalent to R."

The accurate operational semantics of these copulas are provided by the inference rules of NAL [103]; the first two specify when a term can replace another one by meaning, while the last two specify when a statement can replace another one by truth-value. Since terms (including statements as a special type) name concepts, reasoning in NARS is effectively all about the substitutability between concepts and the propagation of such substitutability.

Using the copulas as basic built-in relations, other relations can be defined. For instance, an arbitrary relation R among arguments A, B, and C can be represented in NARS as $(*,\ \mathsf{A},\ \mathsf{B},\ \mathsf{C}) \rightarrow \mathsf{R}$ where the left side of the arrow is a compound term intuitively similar to the Cartesian product defined in set theory, and this statement literally states that the relation among A, B, and C is a special type of the relation R. In this way, the inference rules of NARS only

[26] In AERA, the network is implicit and extremely large, so should not be processed as a graph. In NARS, the network is explicit and can also be huge, though it will never be processed as a whole. In both systems the networks are always processed partly and incrementally. The memory of NARS may look like a semantic network or knowledge graph, yet there are fundamental differences (as explained below).

[27] A concrete example is given latter, where relation R among arguments A, B, and C is represented in NARS as $(*,\ \mathsf{A},\ \mathsf{B},\ \mathsf{C}) \rightarrow \mathrm{R}$ where R is a (relational) concept with learned meaning, while '$\rightarrow$' is a basic relation with built-in meaning.

need to directly process the copulas, as all the other relations can be converted to them equivalently.

The knowledge structure of an AERA agent can be seen as a graph of graphs (hypergraph)—sets of linked nodes that hold *explicit patterns*, and the majority of links are latent (implicit), at any point in time, and only those considered potentially relevant at any point in time are made explicit on demand, at runtime, through active comparison and pattern matching, as mentioned above. This is both how concepts are used and formed over time in AERA. Generally speaking, each node (e.g. a CRM or part of it) may contain (small sets of) variables or terms, some with particular values or ranges, but they mostly also reference other nodes, forming a hierarchy. Concepts, in this approach, are thus associated (both loosely and strongly) *subsets* of the full network, where most of the associations/relations are, again, latent, only to be created upon their use for a particular purpose in particular contexts, through said active pattern matching.

Like in NARS, the knowledge in AERA is experience-grounded, and learning, reasoning, planning, and sensing, are always-on continually-running processes. The vast majority of knowledge in an AERA-based system, including causal-relational models (CRM), are abstraction-neutral, and datatype-neutral and defeasible information structures that directly support deduction and abduction. Deduction in AERA is used for (empirical, defeasible) prediction; abduction is used for producing (evidence-based) explanations and (empirical, defeasible) plans.[28] All of these participate in the overt and covert cognitive operations, operating over concept-based knowledge graphs that are generated on demand, as already mentioned. This then collectively defines their (operational) meaning: Overt operations primarily for controlling interaction with the world in which an AERA agent lives, covert operations for steering how its knowledge grows and its cognition develops over time.

The experience-grounded semantics of NARS and AERA can be captured by the following:

1. Statements and information structures are true to a degree, the truth-value indicating the amount of positive and negative evidence collected for a statement or structure via the system's experience.
2. Whether a concept contains a particular instance (example) or property (attribute) is a matter of circumstantial definition (in NARS it depends on the numerical truth-value mentioned previously; in AERA, active deduction and abduction processes).
3. A conceptual relation R can be specified using compound terms (like Cartesian product, as shown in the previous example in NARS on page 28).
4. A compound term and its components are related by the definition of its compound elements, deriving its meaning from these and its active context (relation to other compound information structures at a particular time).

[28] Induction primarily enters into the picture in the creation of new concepts, although AERA can also use it during knowledge compaction (compression). Ongoing work involves the use of induction also in dynamic analogy-making [78].

5. Events experienced by the system can be temporally or spatially related to each other, and these relations also construct compound terms or information structures.
6. The meaning of a term involved in sensorimotor operations is defined partially by its operation, which is procedural by nature. Executable operations have an associated procedural meaning that reveals what they *do;* these operations can be innate or acquired.

According to CCKR, the meaning of a concept changes over time, not only because the system usually doesn't have the time to consider the "full meaning" of every concept when using it, but also because most of the conceptual relations are irrelevant to the tasks the system is working on at any point in time. For instance, a concept C may be linked with 1000 other concepts at a particular moment, those 1000 links (relations) is what C "means" to the system in general at that moment. However, when processing a task (say, answering a question), only 3 out of the 1000 may actually be used, which defines its "meaning for the current purpose," C'. This C' is "constructed on the fly," i.e. composed from the current context, without a specific pointer or name (e.g. a word used to reference it in language), though it is not really a brand new concept, but a sense (usage) of the existing concept C.

As much of the knowledge in the memory will not be relevant at a particular point in time, it is usually the role of resource control/attention mechanism to select the relevant knowledge for processing at any point in time. For instance, for the purpose of comparing a dog's leg to a human leg, features such as "holding up its body," "propelling body forward by pushing," etc. would be selected as points of such comparison, forming a relational graph for the two legs, yielding differences and similarities.

Since the system does not always know how to approach a new situation, rather than rely on existing algorithms, it must use its available knowledge in a case-by-case manner. A system usually cannot afford the resources to exhaustively explore every possible solution to every new problem or situation, due to a general limitation of time and energy, therefore, it must base its actions (mental and physical) on what seems the most relevant relations of its most relevant concepts, at any point in time [99, 84], which may necessitate relying on analogies and generalizations.

In NARS, all the knowledge about a particular concept, i.e., all of its links to other concepts, are stored in a probabilistic priority queue. When a concept is used, such as in the process of answering a question, these links are considered one by one to see if it leads to a candidate answer. Each time a link is requested, every existing one has a chance to be chosen, depending on its *quality*, *usefulness* in history, and *relevance* to the current context.

Since these contextual factors may change each time the same concept is accessed and used, the pieces of knowledge in it may be retrieved in a different order, and the amount of total knowledge retrieved for a given task depends on the available time, which also changes from situation to situation. Consequently, the "current meaning" of a concept may be more or less different from its "general meaning." When a concept becomes relatively stable (after having been used many times in similar ways), a smaller number of its relations may usually be involved when it is used. In such a case, these relations could be considered its "essence" or even "definition."

Unlike traditional concept ontologies in AI, many of which draw a fully connected knowledge graph by putting permanent links between shared nodes (and most often created by humans), a knowledge base in AERA creates such node linkage – under a particular context at a particular point in time – via matching *candidate* shared nodes (other nodes with shared properties); if a match is created, it becomes a temporarily-shared node of two temporary links in a temporary sub-graph.[29] For a particular context, some nodes may thus be considered (temporarily) related, while under other circumstances they would not. The key reason for this approach is to more flexibly handle novelty: A difference in how matching, and thus linking, is done enables adaptation of existing knowledge to an agent's present (novel) situation and active goals, which may be different from anything that the agent has encountered in unpredictable ways. Thus, most "graphs" in an AERA knowledge base are not, strictly speaking, "existing" so much as latent and "emergent;" any AERA knowledge base supports numerous graphs for numerous purposes through such dynamic node (and link) matching, on-demand. The meaning is therefore also emergent, depending on the current context and the results of this dynamic graph generation through the appropriate matching processes.

4.3 The Creation of a Concept in CCKR

A major difference between CCKR and the traditional concept-level representation approaches is the assumption that concepts are created from scratch by the system, based on its experience, using its existing knowledge and processes. A direct result of this, and another important difference, is that rather than being predetermined by its designer or trained to converge to a stable structure from the outset, the conceptual network in CCKR is adaptive and constantly changing throughout the system's life time. As the system runs, new concepts and conceptual relations are created constantly and old ones may be removed, to free up resources.

How are *new* concepts created? The typical process of constituting concepts proceeds by assembling (associating) small pieces of information (e.g. data coming from sensors and prior knowledge related to these) to improve the system's performance in various cognitive purposes. For instance, the concept of "drinking from a cup" could be assembled out of existing re-purposed knowledge as "pouring liquid into a mouth," and might include the cup's curvature, how it may be grasped and moved, and how it holds liquid that pours out in the direction it's tilted, as required by the shape of the mouth of the drinker.

[29] Note that these are actually implemented through instantiation and binding of templates; thus, a trace of such events can be stored to support later processing, for instance episodic memory.

In particular, new concepts in NARS come into being in three ways:

1. *Novel external experience.* Initially, the memory of NARS only contains a small number of concepts used in the innate operations. Since NARS is always open to input information expressible in Narsese or perceivable by the sensors, new information constantly enters the system's experience, including novel terms and patterns. If such an item can pass the initial competition and enter the memory, a new concept will be created for it.
2. *Constructions of the system.* NARS has a set of inference rules to organize the experience efficiently. For example, when the system notices a light that is red, it will coin a "red light" concept to combine this two pieces of information into one.
3. *Meaning-drift* of existing concepts. Besides the changing topological structure of the concept network, the numerical features (such as truth-values and priority-values) of the vertices and edges are also adjusted from time to time. In the system's life cycle, the state of the concept network never repeats. Since the meaning of a concept is determined by its relations with other concepts, all these changes will more or less change the meaning of the concepts. When the meaning of a concept has changed significantly, very often it is more natural to be taken as a different (new) concept. This kind of *meaning drifting* is gradual and continuous, so it is hard to decide the exact moment when one concept becomes another. Even so, this phenomenon is inevitable in CCKR, which produce both desired results (e.g., adaptation) and undesired ones (e.g., inconsistency).

As the future is unknown, NARS never knows for sure how much each new concept can contribute to its future activities. Consequently, concepts compete for the limited resources (processing time and storage space). As the system matures, the number of concepts in the system and the number of relations in each concept will be roughly constant, bounded by the system's storage and processing capacity.

In AERA, concepts are not contained in static graphs, as mentioned earlier, but constituted on-demand through a dynamic construction process, in light of the agent's active goals and situation. The rules for assembling the concept graphs use (low-level) pattern matching and reasoning. In this approach, concepts are an *emergent second-order* phenomenon. Introspection on how concepts are constructed – i.e. processes that regulate the second-order concept construction – are third-order processes, corresponding (roughly) to what has been called "cognitive development."

The mechanisms for graph creation in AERA, and thus concept formation, include **contextual association** by *similarity*, *salience*, *urgency*, and *usefulness*. The strength of association between any particular subsets of knowledge, at any point in time, is then a product of these four dimensions. The product may itself be context-dependent, but this puts a burden on the bootstrapping of the learning process (unless the agent has a teacher that can help with the bootstrapping), which depends on an initial seed [84] (i.e. the program that initially gets the learning going).[30]

In both NARS and AERA, this approach makes concepts (a) refinable over time, as experience grows (for instance, how a particular word is used), (b) adapted to the current situation and use, and (c) naturally combined with other relevant concepts, for any purpose, on demand (cf. winged basketballs). This also has the benefit that the cognitive mechanisms used for creating concepts are the very same that use them, avoiding infinite regression of machinery for generating meaning.[31]

In AERA, concept assembly may be initiated from a goal whose end state (e.g., quenching one's thirst) is related to an elements presence (e.g. a cup's presence) in the sensory field: Sensed patterns indicating the shape and size of an object (such as a cup, and the liquid it holds), brings relevant knowledge to bear on the goal (via association); predicted goal achievement enables the first steps of a plan (e.g. reaching for the cup) to be produced and committed to. This exact process is also how brand-new concepts are in fact created: By associating patterns with other patterns. We hypothesize that the flexibility with which a cognitive architecture can do this, and the arbitrariness to which it can be done, determines to a significant extent its capacity for abstract thought.

But if concepts are always created on the fly, how could they reach any stability, and the learner employ them reliably for thinking? How could polymorphic and complex – but useful – concepts, like "water" and "living beings," achieve concreteness and stability in such an approach? The answer lies in the emergent interaction between the on-line/on-demand assembly process, the agent's (dynamic) task-environment, and the fine-grain constituents of the agent's knowledge (low-level representation). The persistence or solidity of a concept is determined by the strength of latent associations between its constituents, which is to an important extent determined by permanent regularities in the learner's world; the association can be temporally local (as, for instance, the concept of a 'winged basketball') or deeply entrenched (e.g., the concept of a human face, parts and all).

In the CCKR approach, concepts are thus not only constructed incrementally and piece-wise by the learning system over time, based on experience, they are also more or less re-constructed every time they are used. As the association between a particular situational factors and particular sets of knowledge becomes stronger, due to repeated experiential evidence of their usefulness (among other

[30] The greater the uncertainty about the task-environment for an agent at "birth," the more general-purpose the initial bootstrapping learning programs must be, at the cost of learning time (the larger the number of options one is open to, more considerations are required before you can know what to pay attention to, and thus the longer it takes to start learning). Thus, the more of the agent's learning machinery that is supposed to adapt to circumstances, i.e. learned, the longer it takes to stabilize.

[31] While it is tempting to go into further detail on this particular topic, it involves a number of runtime and implementation issues that are unfortunately too involved to be given a treatment here.

things), the more likely it is that mostly the *same* elemental knowledge will be used for creating *similar* graphs in *similar* situations. Situations with an overlap – especially situations that differ wildly, yet associate with a small subset of the same knowledge – will help highlight that particular subset of knowledge as being relevant across otherwise different situations. This has the desired side-effect of *concept solidification.* Furthermore, the knowledge thus created will become increasingly useful for particular purposes in those situations over time, because it is constantly being improved through experience, and this in turn incrementally increases its conceptual contextualized solidification.[32]

Learning proceeds, in the CCKR framework, by explicitly modeling[33] selected subsets of ongoing experience, resulting in information structures that can (directly or indirectly) inform a learner's future behavior, both overt or covert. The adequacy of any information structure thus created, in light of a particular goal, c.f. "knowledge X can be used for achieving goal Y," must take into consideration the context in which it is to be applied, i.e. "in situation Z".[34] This means that any learning controller that achieves autonomy through the use of explicit information structures in a particular environment – that is, freedom from going back to the lab for re-programming when facing novelty – must also have adequate means for their creation, modification, continuous improvement, and for evaluating their usefulness for various purposes in various contexts. This last part, in our approach, should be achieved through a recursive application of *the same concept-centered mechanisms* being thus evaluated, to avoid infinite growth of the information storage.[35]

[32] We hypothesize that this is in fact where the power of language to shape cognition comes from: It is, among other things, a "concept solidifier" mechanism—making certain concepts more useful than others, as evidenced through communicative experience. It may also be the source of some confusion with a hypothesized "language of thought" [21], which in our view is most aptly thought of not as a language but as a reasoning control system.

[33] The term 'model' is used here in the most general sense, as "the strategically chosen features, functions, and relationships in the learner's (overt and covert) experience, purposefully reformulated as information structures" that are useful for achieving an agent's implicit and explicit goals.

[34] This matches the general definition of a *model*, that is, a model with a referent (whether we assume the referent being something out in the physical world or simply the experience of the learner). For instance, a sowing thread may not be useful as a model for answering questions about the strength of a tree trunk, but it might work for answering whether the tree is long enough to bridge the banks of a river).

[35] One critique of this might be that we are replacing infinite growth with infinite recursion, which would in fact be correct. Of these two evils we prefer to take on the latter challenge rather than the former, because (a) it seems to us that humans can apply their concept-creation abilities to virtually anything, a seemingly important feature of human cognition, and, (b) while humans may be able to create apparently infinite recursions of the kind "I know that you know that I know that you know that I know...," they seem to be quite capable of avoiding it becoming an impediment to organized thought.

It may be useful here to sidestep ever so quickly into human cognition with respect to language. Explaining the role of language in concept use and thought is still a largely unsolved question of human intelligence. In CCKR the mechanisms of language creation and use are directly based on the concept creation and use we have described here. In this view, a word is a particular pattern[36] (whether written or spoken) that can be associated with a particular knowledge subset, through the mechanisms outlined above. The result is that any useful association between that pattern and a set of information structures will solidify over time, upon repeated use and experiential evidence of practicality. Importantly, this solidification is *local* in that due to the compositionality of concepts, only elements hypothesized to be *strictly relevant* to the goals and situation at hand are consolidated, as are subsets of the concepts thus created (i.e. concepts are not "blobs" but rather, "Lego sets"). Context plays an important role here: In AERA the knowledge deemed relevant upon reading or hearing a word will depend on the current situation (determined by input from the sensors) and the currently active goals, controlled by the system's resource management mechanisms. Due to the way its knowledge is encoded and its acquisition proceeds, to teach AERA to use words – and language – is to use them in context. Once learned, AERA can use not only appropriate words but also appropriate syntactic patterns to describe or talk about any of its known concepts, treating the words, e.g., "table," as a pattern whose relationship to models of the concept 'table' is described by *models of the* ***use*** *of the word* "table." As the same symbol (word pattern) is subject to the same knowledge management mechanisms as any other knowledge, procedural or otherwise, a word like "car" will naturally result in slightly different graph construction depending on whether the context is the 19th century or the 21st century. With training, words become useful for thinking, because of their rich associations to other knowledge—and they recall contextually relevant knowledge, just as the sight of a cup will recall concepts related to holding cups, drinking from cups, etc., if your goal is to find a cup, and only if the cup is green when you're looking for "green things."

4.4 CCKR: Properties & Implications

A concept-based cognitive general learner in a complex world creates (and improves) its knowledge piecemeal, from small atomic ("peewee") elements, over time. Acquired knowledge is thus made from components (sub-parts) whose meaning is defined by, and emerges from, their use in bridging between sensation and action, prediction and plan, explanation and problem solving, abstraction and generalization, and so on. In the process of their application for these purposes, the concepts and their relationship with other concepts is reinforced (and re-instated) through the existent relations between their constituent components. Seen this way, a concept is a piece of segmented and compressed multi-faceted model of experience.

[36] To be more specific, a word rightfully forms a *family* of patterns, due to the tolerance for a human language user to its variations in practical scenarios.

A concept in CCKR is thus a collection of information structures that are internally related, and describe multiple aspects of a "semi-unified whole" of closely-related things that are relevant to a being's cognition, including its active goals and proximal immediate surroundings and actions. The concept of a 'table' includes (latent) models of its physical features, what it's for, how it's constructed, etc. It includes all the invariant information that can be abstracted from a set of (experienced and hypothesized) instances of tables (i.e. where the label "table" is appropriate, based in part on social convention). Edge cases test these conventions and the information structures that *produce* the concept, so that's why people typically have to *think* more when asked whether something that's made from unusual materials or is missing some standard features is a an instance of that concept. For instance, is a waist-high flat rock in the forest a table? That depends – in part – on the role of the rock in your activities: the very act of deciding whether something "is a table" (and thus also whether to call it, use it, or treat it as a table) is handled by similar mechanisms that enable the creation of new concepts.

The knowledge contained in a CCKR database can be inspectable and manipulable at a fine-grain "atomic" level[37] because the relations between parts of the knowledge are constructive—i.e. they have a decomposable structure whose parts can be inspected – reflected on – by the system itself. In a given conceptual hierarchy, at the lowest level of any concept are its atomic peewee elements – these cannot be dissected because they don't contain anything smaller-grain. Some atomic elements may be expanded by further learning; for instance, even if the wings of an actual airplane are no more detailed in our mind than those of a paper airplane, we may extend it further when learning that there are flaps at the back edge that can move and internal tanks for holding fuel. Other atomic elements may terminate in some fundamental measurement that can be provided by a controller's transducers, and thus usually pertain to an agent's I/O devices. These are key in grounding all learned concepts, as they play a fundamental role in bootstrapping the agent's learning process itself.

A concept, in this conceptualization, can be a compound information structure composed of other concepts, patterns or partial concepts – that may be manipulated as a set, all at once, while also allowing dissection and analysis, according to needs (decided by an agent's current situation and goals). Concepts are thus clusters of introspectable knowledge that contains overlapping graphs that can be dissected down to the smallest atomic pieces of knowledge.

For a large knowledge base, containing thousands or millions of low-level concepts (e.g., how the apparent shape of objects depends on viewing angle, to take an example), the combinatorics and potential to form sub-graphs will

[37] An analogy to physical atoms – not their Platonic indivisible original version – is somewhat useful here: Physical atoms constitute not only something of a coherent whole, they also inherently encapsulate the rules about which they "play" with their surroundings, which atoms they may pair with, and so on. In this sense, like atoms, models in AERA and terms in NARS are made of parts that determine their behavior in the context of other things.

approach infinity. To be flexible – i.e. to be useful in as many situations for as many purposes as possible, supporting a large variety of potential knowledge graphs – elements of such graphs would need to be re-used across various situations (e.g., the ‘supporting’ role of a human leg is similar to the ‘supporting’ role of a table’s leg—and probably partly responsible for them both being called ‘legs.’). For a large knowledge base it is of course prohibitive to pre-compute every possible information structure that might be needed, and in a complex dynamic world, knowing beforehand everything might be needed in the future is also impossible. These are three reasons why our approach relies on dynamic contextual concept (re-)constitution: Dynamic (re-)creation means that the relevant building blocks can be used for thinking, in light of *any* relevant purpose and situation, making knowledge use very contextualizable. This is also why reflection is called for: A cognitive system must be able to inspect its own operation, to the extent necessary to allow for part-wise manipulation, modification, comparison, and evaluation of any subsets of its knowledge.

The latent concepts in a large knowledge base will be highly context-dependent, putting into question the idea that when a concept is used it is the “same” concept as when it was used last time. This is a feature, not a bug: It means that the relevant parts of a concept are summoned for particular intended purposes. Human concepts exhibit this in well-studied edge cases, such as whether a broken flower vase is still a vase (the answer depends in part on what you plan to do with it—are you testing the power of a new kind of glue or do you have some flowers that need to be put in water?). Many examples of fluid and creative examples of usage can be found in [35].

5 CCKR: Summary of Key Features

CCKR concepts bring a computational “economy of thought” [46] – an Occam’s Razor – that allows a learner to use bounded resources more efficiently. This involves (1) simplification, in the form of summaries of complex experiences, (2) identification of similarities (by grouping), and (3) steering of attention (deciding relevance for efficient steering of resource usage).

A concept-based representation is useful for handling unbounded objects and events, by being domain-free and scale-free, meaning that the ability of compositional concepts to represent information isn’t limited to particular domains and they aren’t bound to a particular level of spatiotemporal description.

CCKR information structures contain the necessary representational features to be *autonomously formed and managed* by the learner, *continuously and consistently* over its whole lifetime, on demand. CCKR knowledge is explicit and new concepts can be constructed from existing concepts using clearly defined operators and explicit relations; existing concepts can be split and re-combined, modified, and changed in part or in whole. CCKR information structures naturally represent hierarchies as well as various types of compound concept, even target other CCKR structures in support of reflection. As a result, CCKR representation supports naturally multiple types of reasoning (deduction, abduction, induction, and analogy), and some of them are often considered as rules of learning, too. This means that a CCKR learner can explicitly and systematically compare and contrast its own experience, reason about its own knowledge and cognitive processes, and produce arguments for its actions. This is a fundamental method by which life-long continuous (cumulative) learning can be realised [85].

In principle, the future is different from the past, and every situation is novel. Adaptation means treating anything novel in relation to the familiar, through decomposition, recognizing their partial similarity and ignoring their differences. A concept in our view provides the mechanisms to achieve Piaget's idea of "assimilation" [66] and Hofstadter's "seeing as" [38], allowing for natural decomposition of information gathered over time through experience.

The knowledge a CCKR system obtains is normally from the processing of multiple tasks, so its meaning, usefulness, and significance to the system is not limited to a single task, as that obtained in an end-to-end black-box learning process that aims at the approximation or optimization of a specific function [84, 107].

Since CCKR represents knowledge as concepts and their relations, a learning system using it will produce knowledge that is easier to align with human concepts than most other approaches to knowledge representation, and therefore easier for humans to understand; it also makes it simpler for a learning system to combine its own experience-based knowledge with given human knowledge.

6 Comparison & Discussion

In the following section, CCKR is compared with other approaches of knowledge representation, especially the symbolic and connectionist schools.

6.1 Theoretical Underpinnings & Methodology

In light of the bigger historical picture, research on representation for intelligence is still in the early stages, which means that a large number of important background assumptions and requirements have yet to be addressed by mainstream research, and many have in fact yet to be brought into the foreground [88]. This includes the theoretical underpinnings of the methodologies. Indications that we are in the early stages of research can be seen in the relatively high number of cognitive features that are absent from most current AI systems—it is for instance still not uncommon to see cognitive architectures that are incapable of learning, and very few of them handle time as a first-class citizen. As there exists no accepted broad theory of intelligence, this is perhaps not surprising. Instead of working towards such a theory, however, which requires elucidating all main background assumptions, the research community has addressed its subject matter largely by fragmenting along topical-methodological lines – neural, symbolic, etc. – and researchers in each camp continue to speak to a (relatively small) community with a limited set of shared background assumptions and terminology, which in turn then remain generally obscure in the larger community. This is the current situation, and it could easily go on unchanged for decades due to the complexity and intricacy of the topic.

To make faster progress, a concerted effort to bridge between sub-fields, with common terminology and definitions, must be undertaken. A simple but important step includes stating clearly in research papers the background assumptions and methodological approach followed. It may be argued that connectionist approaches to AI are not based on any particular theory of intelligence or cognition at all but rather, on an extensive analogy—that "brains are the machinery" of intelligence, so by copying (or rather, taking inspiration from) a subset of its features we can re-create intelligence in a machine. While this makes perfect sense for studying the fine details of brain operation, such as neurons and neural clusters, it is not very useful for understanding the higher levels of organization necessary for cognition—and certainly not the level of human knowledge and concepts. When studying human cognition, *minds* and *brains* are certainly two aspects of a particular incarnation of the same phenomenon, but their relationship is not such that one can be naturally and easily derived from the other. A mapping between these two levels ought to be possible, and will undoubtedly be achieved eventually, but it is far from straightforward. Without a theory of intelligence, it is quite a conundrum which of the thousands – or millions – of biological feature candidates in a working brain should go into a successful implementation of an artificial mind.

It may further be argued that the symbolic approach in AI rests on essentially the same exact theoretical basis as that commonly accepted for computer science. This might not be surprising, considering that historically the origins of the basic ideas on which both are based trace back to approximately the same period in history, when practical electronics, information science, and software development were forming, shortly before and around the middle of last century. "The representations used in computation are symbols and the mind is a computation—the human mind must thus also operate on symbols." No wonder that the founding fathers of AI should think that computers would be able to demonstrate human-level cognition within a decade! But now we know they didn't—not by a long shot.

If current theories of computation suffice to describe computation fully, both practically and theoretically, a theory of cognition cannot be the same as that of computation, whether it be the particulars of the information architecture upon which cognition rests – like is commonly thought – or something else, because they are fundamentally different; the principles allowing human thought to surpass contemporary theories of computation in fundamental ways are still missing. But what could the principles of cognition be, beyond those which describe computers?

Our CCKR approach rests on a theoretical basis for (artificial) intelligence that has some overarching principles to offer that go beyond current theories. It rests on a working definition of intelligence that assumes intelligence fundamentally involves adaptation under insufficient knowledge and resources [104, 83]. It follows from this stance that intelligence operates under *relative rationality* [101], and that intelligent agents acquire their own knowledge through *active construction* of information structures [82]. While the latter point certainly echoes the work of some older psychological theories and schools of thought, our work takes several important steps that, unlike this early work, allows the specification of building instructions for machines that capture these essential properties.

In narrow AI, as in traditional software development, the meaning of anything that an artificial system does can only be interpreted from the point of view of the human creator of that system. This is one key reason why people are often reluctant to attribute understanding or intent to, for instance, AI systems like self-driving cars; in the words of the proverbial man-on-the-street, a self-driving car can't be made responsible for its actions because it doesn't really "know what it's doing." Put differently, its actions don't have any *meaning to itself*—a self-driving car doesn't even have any *handle* on meaning: The context of its actions are not part of its operation. It is a tool, plain and simple, unavoidably resulting from the allonomic engineering approach used in its construction – as with all traditional software development [82].

In contrast to traditional software development methodologies, our approach requires a learning agent to create its own meaning. It does so in the way in that it creates and uses information: The knowledge of an agent with *real* intelligence is not an accurate and objective description of the world, but rather, a particular *encoding of the agent's experience*, with *concepts* as reusable units. The knowledge of such a learning agent is always a *construct*, and its learning thus *constructivist*. It also means that while knowledge is constructed, it is in fact *hypothesized*, because the learner can never know what is "truly out there." This has the important and unavoidable conclusion that knowledge is *never certain* and *all reasoning* that such an agent does, based on this knowledge, is *defeasible* [69, 70] and *non-axiomatic* [93].

Learning, in this view, consists of the construction of interconnected (and partial) models of a learner's accumulated experience, about its own actions, including its own cognitive events (thinking about thinking), and those of the world, in a "language of thought" that supports recursion (i.e., one that is compatible with other knowledge representation and mechanisms). If there is regularity in how the agent's sensors respond to physical forces in the world, the models will reflect features of the agent's environment, as sampled through experience, in a compressed form (which depends in part on the agent's cognitive apparatus; the other part it depends on is the form of the agent's sensors and history of interaction with the environment). Over time this knowledge becomes increasingly useful for meeting the agent's numerous active goal(s). The above considerations form a foundation that separates, in our view, intelligence found in nature from current narrow AI, whether symbolic, sub-symbolic, or something else. They result in a set of implications that, again in our view, makes it clear that current approaches to AI are unlikely to lead to AGI or "real AI." Attempting to put these in a comprehensive list is thus our next task.

6.2 CCKR Concepts vs. Symbols

A predictable misunderstanding of what we have presented is to take CCKR as yet another symbolic approach, since a concept may look a lot like a symbol, which also can serve as an identifier of information structures within a system.

We insist that the concepts in CCKR are not symbols, mainly because of the fundamental difference in how they work, and what a concept and a symbol each *mean*. In the view of CCKR, symbols do exist, but not in the mind: Instead, they are the physically manifested representation of a concept, thought, or relationship – or a sequence thereof – such as a word (composed of letters), a "no-smoking" sign (composed of symbols), a particular temporal sequence of sound waves (speech), etc. Such physical symbols are used for communication to allow the creation of a shared set of concepts, relations, etc., through sensory apparati—physical symbols are manifestations of thought based on concepts.[38]

Further, if 'symbolic' were to be interpreted as equivalent to "using identifiers," the notion would apply to all systems implemented in digital computers, including the connectionist models, as the nodes and links are addressed using internal identifiers or names. So this is clearly inadequate, and a more thorough dissection is called for.

In their "physical symbol system hypothesis," Newell and Simon [56] stated that a symbol "designates" an external object or a process, and the symbol can be "interpreted" according to this mapping. Given this relation, the system's manipulation of the symbol will make it behave according to the object outside the system. This usage of "symbol" follows the referential view of semantics [63], in accordance with traditional *model-theoretic semantics* [6], a well-established theory. This semantics has played a central role in mathematics for specifying the meaning of mathematical concepts, but its application in AI has significant shortcomings. As a result, it has received its share of criticism such as Searle's "Chinese Room" thought experiment [76], which rightfully highlights a major blind spot of all symbolic (GOFAI) AI approaches in that a system using symbols according to model-theoretic semantics has no access to the interpretation of the symbols in it, because according to that theory, the meaning of the symbols is determined by an interpretation that is independent of the system's experience, or the way in which the symbols are processed by the system. The issue is often expressed as the "symbol grounding problem," i.e., the question *"how can the semantic interpretation of a formal symbol system be made intrinsic to the system?"* [29, p. 335].

CCKR is not a solution to this problem but a rejection of its presumptions – that is, of the very foundations of model-theoretic semantics – which define the meaning of a symbol by its denotation outside the system. In CCKR, the meaning of a concept is determined by its role in the system's experience, as revealed by its relations with other concepts [95] and its use for overt and covert action [84]. For this to be possible, the creation, manipulation, and management

[38] The creation, manipulation, comparison, etc. of concepts is, however, "symbolic" in the sense of "using identifiers;" see following paragraph.

of concepts must be accessible to the system and automated in the system's operation, i.e., their meaning must be determined by the system itself at runtime. In this view there is no designation or interpretation from an observer. As long as a concept appears in (or can be related to) the system's experience, it has some (latent[39]) meaning to the system. Here the "experience" can be anything represented as internal events—linguistic, sensorimotor, or anything else that the system experiences, including covert operations, and the relation can be direct or indirect, i.e., via zero or multiple abstractions from the raw input and internal operations.

Specifying the meaning of a concept or a word by its relation to other concepts/words is not a new idea, as proposed in *conceptual role semantics* [28] and the *language-game* theory [111]. In computer science and AI, the approach has taken various forms, from semantic network [112] to knowledge graph [19]. These "relation-based" approaches have mockingly been termed "Dictionary-Go-Round" and criticized as defining a meaningless symbol using other meaningless symbols. This is not the case in CCKR, because here the conceptual relations are not definitions, but rather *experienced* by the system itself—we could perhaps call it "contextuo-temporally grounded" because in this view any experience is couched in the context of internal operations and similar experiences, temporally (and spatially) demarcated.

CCKR requires the concepts to be (eventually) related by the built-in relations that are directly processed by the system, and thus have an in-born operational semantics (seed-based knowledge in AERA; copulas in NARS). The built-in relations do not require any interpretation to become meaningful, and neither do they allow multiple interpretations, as they are used according to the meta-level knowledge generation processes innate to the system [97, 84]. These (innate) concepts have fixed meaning that is not influenced by the system's experience, and they also contribute in determining the meaning of the other (acquired) concepts using the system's experience.

Another common criticism to the relation-based approaches is that a concept can be related to too many other concepts for such a representation to be practical. This issue is resolved in CCKR by its acknowledgment of resource restrictions and dependence on an attention-allocation mechanism, which limits selected relations at the point that each concept is used to the relevant ones for an immediate purpose (temporally defined by the situation and active goals). This attention mechanism decides the current meaning among the general meaning of a concept, and therefore also explains the context-sensitivity of concept usage. This will be controlled by various cognitive operations, many being some form of reasoning, will determine the particulars of how this proceeds.

An important property of CCKR is its dynamic nature. As the conceptual network comes from the system's own experience, it will continue to change

[39] The meaning is latent because one or more such meanings may not be realized for the purposes of a particular situation at a particular time—for this reason it might be better to say that 'a particular concept may have many latent meanings, and specific meaning(s) depending on its use in a particular situation.'

as long as the system is running. In contrast, traditional symbolic AI systems typically assume that every symbol has a fixed meaning that is given by its (fixed) interpretation, being independent of the system's knowledge, history, active goals, and current status.

The claim here is not that model-theoretic semantics is *completely* incompatible with CCKR,[40] but rather, that it should not be applied to the acquired and empirical concepts [82, 98]. The most important similarity of CCKR and symbolic AI is at their level of abstraction, that is, trying to build a *mind*, rather than a *brain*. However, they make very different assumptions about how a mind is related to its environment and what knowledge is: CCKR models the system's interaction with the world, rather than the world "as it really is."

6.3 CCKR conceptual networks vs. artificial neural networks

At a high level of abstraction, CCKR and artificial neural networks (ANNs) share the common features of being networks (graphs) consisting of interconnected units, where the knowledge is distributed in the (semi-explicit) edges, which have "weight values" that can be learned. When the system runs, there is a form of "activation spreading" along these edges. This is mostly where the similarities end, however; CCKR is otherwise very different from ANNs and other connectionist models in many important aspects.

CCKR and ANNs correspond to very different levels of description of the brain, that is, the (psychological) *conceptual* level and the (biological) *neural* level, respectively. A widely-used argument in favor of the ANN approach is that the brain is a neural network, so this is proof that a network-based approach is right for creating intelligence (cf. [77]). The statement is severely misleading, however, in that even though the brain could be fully described as a neural network (at one level of abstraction), this network would still be very different from current ANNs, and it does not exhaust all aspects of the brain functions – a brain is not *merely* a neural network. Furthermore, the validity of this description level does not exclude others. In particular, human introspective experience of our own thinking process is solely at the concept level (we say "I get this idea," not "I have this neural cluster activated"). This is one of the fundamental reasons why the ANN approach to AI runs into trouble when a knowledge-level explanation of their processes is demanded, because we do not understand neural activities (in either their biological forms or their mathematical imitations) at the level where our own learning, reasoning, and thinking are usually explained.

It can be argued that these two levels are related but that the neural level is in a sense "more fundamental." Even if this were the case (which we find a tenuous claim), it does not mean that conceptual-level descriptions have to be reduced to the neural level to be meaningful or useful, or even correct. On the contrary, conceptual level descriptions have the advantage of being independent

[40] There can still be symbols whose meaning is decided by an interpretation, as those in seed-models in AERA, and those used in the axiomatic subsystems in NARS.

to the biological or evolutionary aspects of the human brain, which are not necessarily needed for intelligence in general. Few would argue that the operations of a word processing program should be reduced to the transistor level to be understandable or meaningful. Intelligence is an information process, and any theory of it must address this level at a minimum [55]; bridging to lower levels – explaining how they can be implemented – is an added bonus. A historical reason for why many people prefer neural networks as the theoretical underpinnings of their work is because they see it as a promising alternative to the symbolic representation, which historically has been rigid, fragile, and dependent on human construction [34, 37, 80], and its resemblance to human brain is secondary. Based exclusively on the symbolic AI work of the 60s and 70s, such a view is understandable, but is an unfortunate results of incorrect induction from limited historic events.

Some researchers may argue that there are also 'concepts' in an ANN. Though in a sense the neurons at the input and output layers do share some properties with the concepts in CCKR as "identifiable information structure," they lack the rich semantic, dynamic, and constructive features stressed in CCKR.

Bringing many of the good features of both symbolic AI and ANNs [96], our CCKR approach offers the following desired features:

1. Fluid, adaptive, and flexible ("stretchable") concepts: In CCKR, the boundaries of concepts are not sharp or binary, that is, whether a particular instance of a phenomenon belong to a particular concept is a matter of degree, adjusted by the system itself according to history, purpose, and context.
2. Tolerance to various types of uncertainty: CCKR allows uncertainty in knowledge, so local or global inaccuracies, incompleteness, and inconsistencies will not crush the system.
3. Parallel processing: CCKR allows multiple tasks to be processed concurrently, either in a time-sharing manner or in multiple processors, because it does not require global consistency among the system's beliefs or desires (though the system always tries to maintain these consistencies as much as it can).
4. Self-organization: CCKR supports the construction of new concepts in multilevel abstractions, as feature learning in deep neural networks (DNNs). Unlike DNNs, however, the concepts are made from parts that themselves are individually addressable, inspectable, and expandable, enabling self-organizing compositional concept hierarchies.

These have all been demonstrated in some way in our work on NARS and AERA. Equally importantly, these benefits do not come at a cost but rather, CCKR avoids some major weaknesses exhibited by prior approaches, including ANNs [96]:

- Compositionality: CCKR knowledge is explicit and new concepts can be constructed from existing concepts using clearly defined operators and explicit relations; existing concepts can be split and re-combined, modified, and changed in part or in whole. CCKR information structures naturally represent hierarchies as well as various types of compound concept, even target other CCKR structures in support of reflection.

- Autonomy: CCKR information structures contain the necessary representational features to be *autonomously formed and managed* by the learner, *continuously and consistently* over its whole lifetime, on demand.
- Generality: The knowledge a CCKR system obtains is normally from the processing of multiple tasks, so its meaning, usefulness, and significance to the system is not limited to a single task, as that obtained in an end-to-end black-box learning process that aims at the approximation or optimization of a specific function [84, 107].
- Reasoning: CCKR representation is in large part explicit, and supports naturally many forms of reasoning, including deduction, abduction, induction, and analogy, on any knowledge subset. This means that a CCKR learner can explicitly and systematically compare and contrast its own experience, reason about its own knowledge and cognitive processes, and produce arguments for its actions. This is a fundamental method by which life-long continuous (cumulative) learning can be realised [85].
- Explainability: Since CCKR represents knowledge as concepts and their relations, a learning system using it will produce knowledge that is easier to align with human concepts than most other approaches to knowledge representation, and therefore easier for humans to understand; it also makes it simpler for a learning system to combine its own experience-based knowledge with given human knowledge.

6.4 Unification vs. integration

CCKR is based on the hypothesis that intelligence in general can be fully explained and reproduced at the conceptual level, largely independently of its substrate and underlying lower levels, being it biological neural networks, electrical circuits, or something else. CCKR does not aim at describing human knowledge representation specifically, or in detail, but to abstract its fundamental principles to a level where they are equally applicable to a wide range of computer and natural systems.[41]

While sharing features with both the symbolic and the connectionist approaches, CCKR is neither a hybrid nor an integration of them, but rather a new and broader proposal, resting on a reasoned foundation, that subsumes their best features and overcomes their individual limitations. The resulting CCKR theory does not contain “a symbolic part” and “a connectionist part” but rather, breaks the walls between these two paradigms and gets at the more fundamental principles that their commonalities rest on, achieving a larger, more coherent picture.

[41] Of course, eventually it needs to be explained how such a representation is implemented in the human brain, and how it differs from other animals, but that is neither a topic we consider to be fruitful as a primary methodological principle for AI research in its current state nor one to be addressed in this article.

One fundamental reason for the failure of past attempts at a proper theory of concepts is that many of the strong theoretical principles on which prior attempts have rested are mostly non-existent here, or only present in a weaker form. Without a good (or good enough) theory, attempts at unification ultimately must make do with mere integration instead, where an improper combination inevitably results from incompatible assumptions about the operation of the whole. Yet it is precisely *that which is lacking* that would allow a comprehensive unification: NARS and AERA do not have separate "symbolic modules" and "connectionist modules," nor do they directly use existing logic or ANNs. Instead, they can be considered as the traditional logical or rule-based systems completely rebuilt according to many subsymbolic/connectionist ideas [107, 96, 37, 80].

We agree with many other researchers that a hybrid or integrated approach may be more suitable (in terms of simplicity and efficiency, for instance) for certain specific problems, though it will be difficult, if not impossible, to build a general-purpose intelligent system (AGI) using such an approach. It is true that both paradigms can find supporting evidence in human cognition, as the same cognitive process can be described at either the conceptual level (as in psychology) or the neural level (as in neuroscience), though neither can claim to have captured all the information about the process.

Since connectionist and symbolic approaches correspond to different levels of abstraction when referring to human cognition, each uses a unique vocabulary to explain target phenomena in a somewhat incompatible way. There surely is a correspondence between the neural-level and conceptual-level descriptions of the same process, but it is not a one-to-one mapping; in general, there is no direct mapping between neurons and concepts, nor is there a separate "neural part" and "symbolic part" in the brain. CCKR is 'unified' also because it uses 'concept' in a broad sense to cover the territories which are traditionally referred to as "subsymbolic" or "perceptual." As described previously, since according to CCKR a concept is an internal entity of varying size that can be explicitly identified as a unit, there is no reason to exclude mental image, goal, operation, etc. from forming a concept.

CCKR takes *concept* as a central unit of representation, but does not treat its *meaning* as a constant determined by designation, definition, or third-party interpretation, but rather, being contained in the multifaceted properties learned by the system from experience. This position makes it fundamentally different from other approaches in artificial intelligence and machine learning, though is closer to many research on concepts in cognitive science [43, 35].

In our approach, the meaning of a concept is determined by the system's experience and context, and thus CCKR is also compatible with the view that cognition is *embodied* [2], because a system's "body" is by definition demarcated by sensorimotor operations, which partially constrain the system's experience. When a learning system is implemented with different sensors and actuators, the system's concepts and beliefs will be different, even though the mechanisms responsible for its intelligence (concept-centered representation, unified inference capability, resource allocation procedures, etc.) may remain the same. The changes in body and experience have an impact on the system's problem-solving skills, but not its intelligence, which is defined at the meta-level of the skills, and is (largely) body-independent [100]. Thus, the requirement of embodiment should neither be understood as a requirement for the meaning of concepts to be reduced to the sensorimotor level, nor the requirement that the system's body (and experience) be close to that of a human [4].

By extension, CCKR also covers *situated* and *context-sensitive* cognition [3], because its memory is dynamically structured to favor the concepts and relations that are directly related to the current situation, or the need raised by it, as explained in [97]. However, it does not mean that the system is exclusively reactive, with no long-term motivations and drives; quite the contrary: the CCKR approach fully supports reasoning, widely construed, including prediction, explanation, causal reasoning, and analogy making (though in this article they cannot be described). Here the *context* of a concept is provided by other related concepts and their current role in light of active goals and situations, rather than by explicit labels, as in the traditional symbolic AI approaches [7, 49].

7 Conclusions

Our proposed concept-centered knowledge representation (CCKR) is directly based on our understanding of intelligence, cognition, and mind. While resting on – and echoing – many prior ideas in the field of AI and cognitive science, we consider our formulation new in that it is more specific and offers, and contrasts here, two implementations (NARS and AERA) that have already demonstrated to go beyond other methods in machine learning and AI in important ways. While both are certainly in their early stages, we consider them sufficiently mature to compare to prior attempts at unified theories of cognition. This paper focuses on the information content and form of our proposed concept-centered knowledge representation demonstrated in these systems, but our work extends well beyond these, sufficiently so to allow functionally operational software demonstrations of its key principles.

CCKR does not assume the existence of an external world that consists of clearly separated objects and events waiting for the system to discover, recognize, or manipulate them. On the other hand, this position does not assume an arbitrary or random environment either. We assume that an environment exists independently of the system and is complex enough that its future cannot accurately be predicted in full detail, while containing sufficient regularity for learning through experience, since otherwise no adaptation is possible.

CCKR proposes that in any general-purpose intelligent system, knowledge must be represented as a dynamic conceptual network with particular properties and related operations. This network is neither a "neural" network, as each concept has an internal identifier function that allows the network to be used to construct more complicated concepts, to be constructed, re-constructed and de-constructed, and is therefore not restricted to a specific task or function. Nor is the network a "symbolic" network as classically construed, since the concepts do not refer to external objects and events; the network's structure, as well as the concepts and their relations, constantly change as the system interacts with the environment while accomplishing its tasks through continual reasoning and learning.

CCKR is presented as a set of principles that allow different formalization and implementation, rather than a single model. This is the same case for the symbolic and connectionist approaches, as both of them include various concrete designs, while still having enough features to be distinguished from each other. For this reason, we have not forced CCKR into a single formal model, though it still shows enough differences from the other two.

A concept in our approach is neither a symbol, nor does it necessarily represent an external object or event, but rather, an ingredient, pattern, and model of the system's *partial* experience, that is, a record of its interaction with the environment and a record of its own use of its knowledge. While some basic concepts can be innate or given, most concepts are constructed by the system itself, and all concepts are modifiable by the system according to history and context, and with different degrees of stability and clarity.

The benefits of a concept-centered knowledge representation include:

- Treating novelty through decomposition and partial similarity to what is already known, ignoring and/or examining the differences. This ability is fundamental to all intelligence because the future is always different from the past and every situation is novel.
- More efficient use of resources in representing and learning from experience, bring a 'computational economy of thought:' Concepts allow for (1) simplification, in the form of summaries of detailed, complex experiences, (2) identification of similarities (by grouping), and (3) steering of attention (using estimates of relevance to achieve more efficient steering of resource usage).
- Freedom from predetermined levels of spatiotemporal descriptions: A concept-based representation is domain-free and scale-free, meaning that the compositional concepts can represent information that isn't limited to particular domains or bound to specific level of spatiotemporal description, allowing them to handle unbounded objects and events.

References

1. Anderson, J.R., Lebiere, C.: The Atomic Components of Thought. Lawrence Erlbaum Associates, Mahwah, New Jersey (1998)
2. Anderson, M.L.: Embodied cognition: a field guide. Artificial Intelligence **149**(1), 91–130 (2003)
3. Barsalou, L.W.: The instability of graded structure: implications for the nature of concepts. In: Neisser, U. (ed.) Concepts and Conceptual Development: Ecological and intellectual factors in categorization, pp. 101–140. Cambridge University Press, Cambridge (1987)
4. Barsalou, L.W.: Perceptual symbol systems. Behavioral and Brain Sciences **22**, 577–609 (1999)

5. Barsalou, L.W.: Challenges & opportunities for grounding cognition. Journal of Cognition **3**, 1–24 (2020)
6. Barwise, J., Etchemendy, J.: Model-theoretic semantics. In: Posner, M.I. (ed.) Foundations of Cognitive Science, pp. 207–243. MIT Press, Cambridge, Massachusetts (1989)
7. Barwise, J., Perry, J.: Situations and Attitudes. MIT Press, Cambridge, Massachusetts (1983)
8. Beer, R.: Dynamical approaches to cognitive science. Trends in Cognitive Sciences **4**, 91–99 (2000)
9. Besold, T.R., Garcez, A., Bader, S., Bowman, H., Domingos, P.M., Hitzler, P., Kühnberger, K.U., Lamb, L., Lowd, D., Lima, P., Penning, L., Pinkas, G., Poon, H., Zaverucha, G.: Neural-symbolic learning and reasoning: A survey and interpretation. ArXiv **abs/1711.03902** (2017)
10. Birnbaum, L.: Rigor mortis: a response to Nilsson's "Logic and artificial intelligence". Artificial Intelligence **47**, 57–77 (1991)
11. Brooks, R.A.: Intelligence without reason. In: Proceedings of the 12th International Joint Conference on Artificial Intelligence. pp. 569–595 (1991)
12. Brooks, R.A.: Intelligence without representation. Artificial Intelligence **47**, 139–159 (1991)
13. Brooks, R.A., Breazeal, C., Irie, R., Kemp, C.C., Marjanovic, M., Scassellati, B., Williamson, M.M.: Alternative essences of intelligence. In: Proceedings of the Fifteenth AAAI/IAAI Conference. pp. 961–968 (1998)
14. Chomsky, N.: Syntactic Structures. Mouton, The Hague (1957)
15. Conant, R.C., Ashby, W.R.: Every good regulator of a system must be a model of that system. International journal of systems science **1**(2), 89–97 (1970)
16. Dinsmore, J. (ed.): The Symbolic and Connectionist Paradigms: Closing the Gap. Lawrence Erlbaum Associates, Inc. (1992)
17. Dong, H., Mao, J., Lin, T., Wang, C., Li, L., Zhou, D.: Neural logic machines. In: Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019) (2019), https://openreview.net/pdf?id=B1xY-hRctX
18. Dreyfus, H.L.: What Computers Can't Do: Revised Edition. Harper and Row, New York (1979)
19. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems - SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics. pp. 13–16 (2016)
20. Ensmenger, N.: Is chess the drosophila of artificial intelligence? A social history of an algorithm. Social Studies of Science **42**(1), 5–30 (2011)
21. Fodor, J.A.: The Language of Thought. Thomas Y. Crowell, New York (1975)
22. Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: a critical analysis. Cognition **28**, 3–71 (1988)
23. d'Avila Garcez, A., Gori, M., Lamb, L.C., Serafini, L., Spranger, M., Tran, S.N.: Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. FLAP **6**(4), 611–632 (2019), https://collegepublications.co.uk/ifcolog/?00033
24. van Gelder, T.: The dynamical hypothesis in cognitive science. Behavioral and Brain Sciences **21**(5), 615–628 (1998)
25. Gomila, T., Calvo, P.: Directions for an embodied cognitive science: Toward an integrated approach. In: Handbook of cognitive science, pp. 1–25. Elsevier (2008)

26. Goschke, T., Koppelberg, D.: Connectionist representation, semantic compositionality, and the instability of concept structure. Psychological Research **52**, 253–270 (1990)
27. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwi?ska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A.P., Hermann, K.M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., Hassabis, D.: Hybrid computing using a neural network with dynamic external memory. Nature **538**(7626), 471–476 (2016), http://dx.doi.org/10.1038/nature20101
28. Harman, G.: Conceptual role semantics. Notre Dame Journal of Formal Logic **28**, 252–256 (1982)
29. Harnad, S.: The symbol grounding problem. Physica D **42**, 335–346 (1990)
30. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**, 245–258 (07 2017). https://doi.org/10.1016/j.neuron.2017.06.011
31. Haugeland, J.: Artificial Intelligence: The Very Idea. Massachusetts Institute of Technology, USA (1985)
32. Hawkins, J., Blakeslee, S.: On Intelligence. Times Books, New York (2004)
33. Hayes, P.J.: The naïve physics manifesto. In: Michie, D. (ed.) Expert Systems in the Micro-Electronic Age, pp. 242–270. Edinburgh University Press, Edinburgh (1979)
34. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representation. In: Rumelhart, D.E., McClelland, J.L. (eds.) Parallel Distributed Processing: Exploration in the Microstructure of cognition, Vol. 1, Foundations, pp. 77–109. MIT Press, Cambridge, Massachusetts (1986)
35. Hofstadter, D., Sander, E.: Surfaces and Essences: Analogy as the Fuel and Fire of Thinking. Basic Books (2013)
36. Hofstadter, D.R.: The copycat project: An experiment in nondeterminism and creative analogies. AI memo, MIT Artificial Intelligence Laboratory (1984)
37. Hofstadter, D.R.: Waking up from the Boolean dream, or, subcognition as computation. In: Metamagical Themas: Questing for the Essence of Mind and Pattern, chap. 26. Basic Books, New York (1985)
38. Hofstadter, D.R.: On seeing A's and seeing As. Stanford Humanities Review **4**, 109–121 (1995)
39. Holland, J.H.: Escaping brittleness: the possibilities of general purpose learning algorithms applied to parallel rule-based systems. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.) Machine Learning: an artificial intelligence approach, vol. II, pp. 593–624. Morgan Kaufmann, Los Altos, California (1986)
40. Kneale, W., Kneale, M.: The development of logic. Clarendon Press, Oxford (1962)
41. Laird, J.E.: The Soar Cognitive Architecture. MIT Press, Cambridge, Massachusetts (2012)
42. Lakoff, G.: Cognitive semantics. In: Eco, U., Santambrogio, M., Violi, P. (eds.) Meaning and Mental Representation, pp. 119–154. Indiana University Press, Bloomington, Indiana (1988)
43. Laurence, S., Margolis, E.: Concepts and cognitive science. In: Margolis, E., Laurence, S. (eds.) Concepts: Core Readings, pp. 3–81. MIT Press, Cambridge, Massachusetts (1999)
44. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015). https://doi.org/10.1038/nature14539

45. Lenat, D.B., Feigenbaum, E.A.: On the thresholds of knowledge. Artificial Intelligence **47**, 185–250 (1991)
46. Mach, E.: The Science of Mechanics. Read Books (2008)
47. Marcus, G.: The next decade in AI: Four steps towards robust artificial intelligence (2020)
48. McCarthy, J.: Artificial intelligence, logic and formalizing common sense. In: Thomason, R.H. (ed.) Philosophical Logic and Artificial Intelligence, pp. 161–190. Kluwer, Dordrecht (1989)
49. McCarthy, J.: Notes on formalizing contexts. In: Proceedings of the thirteenth international joint conference on artificial intelligence. pp. 555–560 (1993)
50. McCulloch, W.S., Pitts, W.H.: A logical calculus of ideas immanent in neural activity. Bulletin of Mathematical Biophysics **5**, 115–133 (1943)
51. McDermott, D.: A critique of pure reason. Computational Intelligence **3**, 151–160 (1987)
52. Minsky, M.: Neural nets and the brain-model problem. Ph.D. thesis, Department of Computer Science, Princeton University (1954)
53. Minsky, M.: A framework for representing knowledge (1974), MIT-AI Laboratory Memo No. 306
54. Minsky, M.: The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. Simon & Schuster (2006)
55. Newell, A.: The knowledge level: Presidential address. AI Magazine **2**(2), 1–20 (1981)
56. Newell, A., Simon, H.A.: Computer science as empirical inquiry: symbols and search. Communications of the ACM **19**(3), 113–126 (1976)
57. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (2015)
58. Nivel, E., Thórisson, K.: Self-programming: Operationalizing autonomy. In: Proceedings of the Second Conference on Artificial General Intelligence. pp. 150–155 (2009)
59. Nivel, E., Thórisson, K.: Towards a programming paradigm for control systems with high levels of existential autonomy. In: Proceedings of the Sixth Conference on Artificial General Intelligence. pp. 78–87 (2013)
60. Nivel, E., Thórisson, K.R., Dindo, H., Pezzulo, G., Rodríguez, M., Hernandez, C., Steunebrink, B.R., Ognibene, D., Chella, A., Schmidhuber, J., Sanz, R., Helgason, H.P.: Autocatalytic Endogenous Reflective Architecture. Reykjavik University School of Computer Science Technical Report, RUTR-SCS13002 (2013), https://alumni.media.mit.edu/~kris/ftp/AERA-RUTR-SCS13002.pdf
61. Nivel, E., Thórisson, K.R., Steunebrink, B., Schmidhuber, J.: Anytime bounded rationality. In: Bieger, J., Goertzel, B., Potapov, A. (eds.) Proceedings of the 8th Conference on Artificial General Intelligence. pp. 121–130. Springer International Publishing, Cham (2015)
62. Nivel, E., Thórisson, K.R., Steunebrink, B.R., Dindo, H., Pezzulo, G., Rodríguez, M., Hernández, C., Ognibene, D., Schmidhuber, J., Sanz, R., Helgason, H.P., Chella, A., Jonsson, G.K.: Bounded recursive self-improvement. CoRR **abs/1312.6764** (2013), http://arxiv.org/abs/1312.6764
63. Ogden, C., Richard, I.: The Meaning of Meaning (10th edition, 1949). Kegan Paul, London (1923)
64. Pattee, H.H.: Cell psychology: An evolutionary approach to the symbol-matter problem. Cogn. Brain Theory **5**, 325–341 (1983)

65. Pearl, J., Mackenzie, D.: The Book of Why. Basic Books, New York (2018)
66. Piaget, J.: The construction of reality in the child. Basic Books, New York (1954)
67. Piccinini, G.: The first computational theory of mind and brain: A close look at McCulloch and Pitts's 'logical calculus of ideas immanent in nervous activity'. Synthese **141**, 175–215 (April 2004)
68. Pinker, S., Prince, A.: On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition **28**, 73–193 (1988)
69. Pollock, J.: OSCAR: an architecture for generally intelligent agents. In: Proceedings of the First Conference on Artificial General Intelligence. pp. 275–286 (2008)
70. Pollock, J.L.: Thinking about Acting: Logical Foundations for Rational Decision Making. Oxford University Press, USA, New York (2006)
71. Rogers, T.T., McClelland, J.L.: Précis of semantic cognition: A parallel distributed processing approach. Behavioral and Brain Sciences **31**, 689–749 (2008)
72. Rosenblatt, F.: The perceptron: A perceiving and recognizing automaton (1957), cornell Aeronautical Laboratory Report 85-60-1
73. Rosenbloom, P.S., Demski, A., Ustun, V.: The Sigma Cognitive Architecture and System: Towards Functionally Elegant Grand Unification. Journal of Artificial General Intelligence **7**(1), 1–103 (2016)
74. Rumelhart, D.E., McClelland, J.L. (eds.): Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, Foundations. MIT Press, Cambridge, Massachusetts (1986)
75. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Networks **61**, 85–117 (2015)
76. Searle, J.: Minds, brains, and programs. Behavioral and Brain Sciences **3**, 417–424 (1980)
77. Sejnowski, T.J.: The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences (2020). https://doi.org/10.1073/pnas.1907373117
78. Sheikhlar, A., Thórisson, K.R.: Causal generalization via goal-driven analogy. In: Thórisson, K.R., Isaev, P., Sheikhlar, A. (eds.) International Conference on Artificial General Intelligence. pp. 165–175. Springer Verlag (2024)
79. Skinner, B.F.: About Behaviorism. Knopf Doubleday Publishing Group, New York (1974)
80. Smolensky, P.: On the proper treatment of connectionism. Behavioral and Brain Sciences **11**, 1–74 (1988)
81. Sun, R.: Robust reasoning: integrating rule-based and similarity-based reasoning. Artificial Intelligence **75**, 241–295 (1995)
82. Thórisson, K.R.: A new constructivist AI: From manual methods to self-constructive systems. In: Wang, P., Goertzel, B. (eds.) Theoretical Foundations of Artificial General Intelligence, pp. 145–171. Atlantis Press, Paris (2012)
83. Thórisson, K.R.: Discretionarily constrained adaptation under insufficient knowledge & resources. Journal of Artificial General Intelligence **11**(2), 7–12 (2020)
84. Thórisson, K.R.: Seed-programmed autonomous general learning. Proceedings of Machine Learning Research **131**, 32–70 (2020)
85. Thórisson, K.R., Bieger, J., Li, X., Wang, P.: Cumulative learning. In: Proceedings of the 12th International Conference on Artificial General Intelligence. pp. 198–208. Springer Verlag (2019)
86. Thórisson, K.R., Helgasson, H.P.: Cognitive architectures and autonomy: A comparative review. Journal of Artificial General Intelligence **3**(2), 1–30 (2012)

87. Thórisson, K.R., Kremelberg, D., Steunebrink, B.R., Nivel, E.: About understanding. In: The Proceedings of the Ninth Conference on Artificial General Intelligence. pp. 106–117 (2016)
88. Thórisson, K.R., Minsky, H.: The future of ai research: Ten defeasible 'axioms of intelligence'. Proc. Machine Learning Research **192**, 5–21 (2023)
89. Thórisson, K.R., Nivel, E., Sanz, R., Wang, P.: Approaches and Assumptions of Self-Programming in Achieving Artificial General Intelligence. Journal of Artificial General Intelligence **3**(3), 1–10 (2012)
90. Thórisson, K.R., Nivel, E., Steunebrink, B.R., Helgason, H.P., Pezzulo, G., Sanz, R., Schmidhuber, J., Dindo, H., Rodriguez, M., Chella, A., Jonsson, G.K., Ognibene, D., Hernandez, C.: Autonomous Acquisition of Situated Natural Communication. Computer Science & Information Systems **9**(2), 115–131 (2014), *Outstanding Paper Award*
91. Thórisson, K.R., Talbot, A.: Cumulative learning with causal-relational models. In: Artificial General Intelligence. pp. 227–237. Springer International Publishing, Cham (2018)
92. Thórisson, K.R., Talevi, G.: A theory of foundational meaning generation in autonomous systems, natural & artificial. In: Thórisson, K.R., Isaev, P., Sheikhlar, A. (eds.) International Conference on Artificial General Intelligence. pp. 188–198. Springer Verlag (2024)
93. Wang, P.: Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. Ph.D. thesis, Indiana University (1995)
94. Wang, P.: The logic of learning. In: Working Notes of the AAAI workshop on New Research Problems for Machine Learning. pp. 37–40. Austin, Texas (2000)
95. Wang, P.: Experience-grounded semantics: a theory for intelligent systems. Cognitive Systems Research **6**(4), 282–302 (2005)
96. Wang, P.: Artificial general intelligence and classical neural network. In: Proceedings of the IEEE International Conference on Granular Computing. Atlanta, Georgia (2006)
97. Wang, P.: Rigid Flexibility: The Logic of Intelligence. Springer, Dordrecht (2006)
98. Wang, P.: Three fundamental misconceptions of artificial intelligence. Journal of Experimental & Theoretical Artificial Intelligence **19**(3), 249–268 (2007)
99. Wang, P.: Case-by-case problem solving. In: Goertzel, B., Hitzler, P., Hutter, M. (eds.) Proceedings of the Second Conference on Artificial General Intelligence. pp. 180–185 (2009)
100. Wang, P.: Embodiment: Does a laptop have a body? In: Goertzel, B., Hitzler, P., Hutter, M. (eds.) Proceedings of the Second Conference on Artificial General Intelligence. pp. 174–179 (2009)
101. Wang, P.: The assumptions on knowledge and resources in models of rationality. International Journal of Machine Consciousness **3**(1), 193–218 (2011)
102. Wang, P.: Motivation management in AGI systems. In: Bach, J., Goertzel, B., Iklé, M. (eds.) Proceedings of the Fifth Conference on Artificial General Intelligence. pp. 352–361 (2012)
103. Wang, P.: Non-Axiomatic Logic: A Model of Intelligent Reasoning. World Scientific, Singapore (2013)
104. Wang, P.: On defining artificial intelligence. Journal of Artificial General Intelligence **10**(2), 1–37 (2019). https://doi.org/10.2478/jagi-2019-0002
105. Wang, P.: Toward a logic of everyday reasoning. In: Vallverdú, J., Müller, V.C. (eds.) Blended Cognition: The Robotic Challenge, pp. 275–302. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-03104-6_11

106. Wang, P., Hofstadter, D.: A logic of categorization. Journal of Experimental & Theoretical Artificial Intelligence **18**(2), 193–213 (2006)
107. Wang, P., Li, X.: Different conceptions of learning: Function approximation vs. self-organization. In: Steunebrink, B., Wang, P., Goertzel, B. (eds.) Proceedings of the Ninth Conference on Artificial General Intelligence. pp. 140–149 (2016)
108. Wang, P., Steunebrink, B., Thórisson, K.R.: What should AGI learn from AI and CogSci? In: Goertzel, B., Orseau, L., Snaider, J. (eds.) Presented at the International Conference on Artificial General Intelligence. Springer Verlag (2014)
109. Watson, J.B.: Psychology as the behaviorist views it. Psychological Review **20**(2), 158–177 (1913)
110. Winograd, T.: Thinking machines: Can there be? Are we? In: The Foundation of Artificial Intelligence—a Sourcebook, pp. 167–189. Cambridge University Press, USA (1990)
111. Wittgenstein, L.: Philosophical Investigations. Prentice Hall, Upper Saddle River, New Jersey (1999), translated by G. Anscombe
112. Woods, W.A.: What's in a link: Foundations for semantic networks. In: Bobrow, D.G., Collins, A.M. (eds.) Representation and Understanding: Studies in Cognitive Science,, pp. 35–82. Academic Press, New York (1975)