IADIS International Journal on Computer Science and Information Systems
Vol. 9, No. 2, pp. 115-131
ISSN: 1646-3692

AUTONOMOUS ACQUISITION OF NATURAL SITUATED COMMUNICATION

Kristinn R. Thórisson. *Reykjavik University / CADIA & Icelandic Institute for Intelligent Machines*.

Eric Nivel. *Icelandic Institute for Intelligent Machines*.

Bas R. Steunebrink. *The Swiss AI Lab IDSIA, USI & SUPSI*.

Helgi P. Helgason. *Icelandic Institute for Intelligent Machines*.

Giovanni Pezzulo. *Consiglio Nazionale delle Ricerche / ISTC*.

Ricardo Sanz. *Universidad Politecnica de Madrid / ASLAB*.

Jürgen Schmidhuber. *The Swiss AI Lab IDSIA, USI & SUPSI*.

Haris Dindo. *Universita degli studi di Palermo / DINFO*.

Manuel Rodriguez. *Universidad Politecnica de Madrid / ASLAB*.

Antonio Chella. *Universita degli studi di Palermo / DINFO*.

Gudberg K. Jonsson. *University of Iceland / Human Behavior Laboratory*.

Dimitri Ognibene. *Department of Informatics, King's College London*.

Carlos Hernandez. *Universidad Politecnica de Madrid / ASLAB*.

ABSTRACT

An important part of human intelligence, both historically and operationally, is our ability to communicate. We learn how to communicate, and maintain our communicative skills, in a society of communicators – a highly effective way to reach and maintain proficiency in this complex skill. Principles that might allow artificial agents to learn language this way are incompletely known at present

– the multi-dimensional nature of socio-communicative skills are beyond every machine learning framework so far proposed. Our work begins to address the challenge of proposing a way for observation-based machine learning of natural language and communication. Our framework can learn complex communicative skills with minimal up-front knowledge. The system learns by incrementally producing predictive models of causal relationships in observed data, guided by goal-inference and reasoning using forward-inverse models. We present results from two experiments where our S1 agent learns human communication by observing two humans interacting in a realtime TV-style interview, using multimodal communicative gesture and situated language to talk about recycling of various materials and objects. S1 can learn multimodal complex language and multimodal communicative acts, a vocabulary of 100 words forming natural sentences with relatively complex sentence structure, including manual deictic reference and anaphora. S1 is seeded only with high-level information about goals of the interviewer and interviewee, and a small ontology; no grammar or other information is provided to S1 a priori. The agent learns the pragmatics, semantics, and syntax of complex utterances spoken and gestures from scratch, by observing the humans compare and contrast the cost and pollution related to recycling aluminum cans, glass bottles, newspaper, plastic, and wood. After 20 hours of observation S1 can perform an unscripted TV interview with a human, in the same style, without making mistakes.

KEYWORDS

Knowledge acquisition, natural language, situated, communication.

1. INTRODUCTION

One of the most useful skills to evolve in humans is the ability to communicate, which serves the function of transferring compressed information between individuals and groups. The skill builds on several co-dependent sub-skills and abilities, such as auditory timbre discrimination, sequence learning, fine motor control, and context-sensitive abstraction, whose evolution and honing over tens of thousands of homo sapiens generations has lead to the diverse use of communication observed in modern human society. The best – and possibly only – way to learn communication for a human is through observation of social interaction, where the effect of language use on oneself and other language users occurs naturally (cf. Petit et al. 2012), with practical needs and constraints driving the learning; where vast numbers of successful and unsuccessful uses of language variations can be related to one's own and others' goals, where numerous exceptions and contextualized cues for usage help define and hone the meaning of concepts and utterances, and where explicit and implicit usage "experiments" of communicative devices can be made directly. If our aim is to create an artificial agent that masters the numerous facets and subtleties of human communication this is probably the case as well: The agent should be situated in some kind of social context, where it can acquire the necessary skills through the same means. This would, however, require a new kind of machine learning, one that could not only observe and imitate what other agents do but that could also penetrate the agents' goals, so that the learning could be derived from a deeper understanding of the agents' intentions, allowing the observation to unlock the methods others use for achieving their goals. While no principles for such a mechanism have been fielded as of yet, our work proposes a way to achieve this.

We present results from experiments with a new type of architecture and methodology aimed at the deep questions of autonomous acquisition of communicative skills. The approach relates closely to other challenges in artificial intelligence, such as life-long learning,

continuous adaptation, and self-programming (Nivel et al. 2013); the focus in this paper, however, centers on communication. Our experiments situate our system in a realtime social interaction similar to a television interview – a domain selected due to its inherent properties of relatively high complexity and real-world constraints. Even more importantly, the multi-layered structure of social interaction – both ontologically and temporally, with sounds forming words, words and gestures forming utterances, utterances forming speech acts, and speech acts and turntaking structures forming social scenarios – includes challenges for which no satisfactory solutions exist at present. Our system learns situated multimodal communication using a single learning mechanism – no separate learning methods, modules, or other exceptions are needed for the various aspects of the scenario, such as gesturing, gesture-speech coordination, word order, question-answer paring, turntaking, and the like.

Based on a new constructivist methodology (Thórisson 2009, 2012) that puts the autonomy of the agent as a main priority, we target systems that can bootstrap their learning from very primitive beginnings. This type of system has the highest potential for adaptation in light of radical changes, both to their own processing resources, their tasks, and their operating environment (cf. Thórisson 2013). For this reason we do away with allonomic[1] approaches to software development, on which all common software development methodologies are currently based (where the human programmer provides a system with its algorithms), and replace it with a self-programming approach in which the majority of the system, upon reaching maturity, consists of code produced by the system itself. At present, an intelligent agent in our framework is provided with a small seed consisting of an object ontology, a handful of top-level goals, and optionally a couple of domain-related goals to help with the bootstrapping (five, in the case of the TV interview). Due to our agents being situated in a social interaction scenario and being engineered to learn continuously, their knowledge is acquired incrementally over time, growing directly and solely from their own experience.

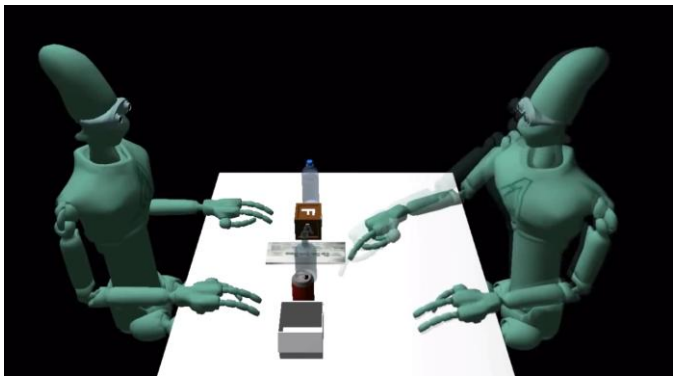

Figure 1. The realtime interaction between a human (interviewer, right) and the S1 agent (interviewee, left), in the form of a simple TV interview conducted in the virtual equivalent of a video conferencing. Live tracking of the human's multimodal behavior and speech directly drives the behavior of the avatar; S1 controls the other avatar via a software API. On the virtual table between the human and S1 are six objects of various materials; the interviewer's role is to get the interviewee to tell about the recycling of these materials, and the comparison of cost, pollution, etc. of creating objects from scratch versus recycling objects of the same and different materials

1 'Allonomy' is the opposite of autonomy; allonomic controllers may impart some level of autonomy to what they control while not being autonomous themselves.

The paper is organized as follows. We first give a short introduction to the theoretical foundation of our work, then we describe the framework we have developed, followed by a description of how this system learns communicative skills by observation. We then describe the two experiments we have run, and the results from these.

2. METHODOLOGICAL FOUNDATION

To construct highly adaptive systems means we must let go of the idea that we, the designers, provide the system with domain- and task-specific algorithms, as this would mean that *we* would have to pick the most important tasks the system is to perform, and proceed to codify by hand all information and relevant algorithms the agent is to follow – clearly a condition in conflict with the goal of autonomy. Our approach therefore requires us to rule out allonomic software methodologies – methodologies that fundamentally rest on hand-coding of domain-level operational functionality. High levels of autonomy means high levels of domain independence, so we also cannot allow ourselves to provide the system mainly with domain-specific knowledge. Instead we focus, in constructivist fashion, on developing general principles to allow the system itself to *invent algorithms*. And we must go even further, for high levels of autonomy means that the system we target must constantly be learning, by training itself on appropriate tasks and subtasks, also after it "leaves the lab". As it turns out, the term "algorithm" may not be entirely appropriate for what our autonomous system is learning, because even on sequential repeats of the same task the system may be modifying how it does it (cf. Wang 2006), from the smallest to the largest subtask. In fact, in our constructivist approach the system development task becomes that of designing a meta-control scheme that, instead of providing hand-coded solutions to specified tasks and subtasks, must give the system enough flexibility and initiative to propose subgoals on its own, based on the drives (highest-level goals) provided by the system's designers, and model its experience in a way that continuously increases its ability to explain (by prediction) both its own behavior and that of its environment.

Our constructivist AI methodology (CAIM) is outlined in Thórisson (2012, 2009), in Nivel & Thórisson (2009) and in Thórisson & Nivel (2009a, 2009b); here we give a quick overview as relates to the present work. CAIM sprang out of two separate threads that are co-related. The first is the view that intelligent beings construct their own knowledge from experience. Sure, inborn principles guide human learning, but evidence suggests that producing effective thinking requires interplay between a mind and its environment. This observation is where Piaget's constructivist proposal originated, and his theory of cognitive stages (Piaget 1950). The second is a view of autonomy that sees intelligence as an extreme form of adaptation capabilities, a view that sees any intelligent system equipped not only with the ability to follow rules, but to *figure out* the rules. Defining mechanisms of human minds, such as the ability to discover, understand and abstract facts and causal chains, to make analogies and inferences, and to learn a large amount of vastly different skills, some of which are historically brand new (cf. space walks, Internet browsing, and flying airplanes), make it clear that providing knowledge up front for these skills takes more than inventing effective algorithms for a few specific tasks – it requires something more general. Paraphrasing Wang's (2004) analogy, to get generality a small set of hand tools won't do, what we need is a *hand*.

The meaning of autonomy is defined in part by the handling of obstacles; greater autonomy means that a better handling of obstacles, in arity, time, and/or complexity. A major source of obstacles for natural intelligences is the complex, nondeterministic nature of the real world, and numerous resource constraints that prevent actors from being able to spend infinite amounts of time thinking. The world has an ever-ticking real-world clock, and energy constraints limit natural computers (brains) to re-structuring themselves into topologies that are supported by the laws of physics. In fact, physical constraints are the reasons for why intelligence arose in the first place. Limitations of computation mean that an embodied agent situated in a complex environment can neither be assumed to process every input available in its environment nor to follow every thought to its ultimate conclusion: Real-world agents do not have the resources to accomplish *all* the jobs they ideally should or could, given their goals, due to limited computing and memory capacity. The assumption of limited resources has fundamental implications for our approach and the design of our auto-catalytic, endogenous, reflective control architecture AERA. Similar to Wang's NARS (Wang 2011, 2006), our approach centers on assumptions of self-bootstrapping from incomplete knowledge and insufficient resources (Thórisson 2013, Nivel et al. 2012, Wang 2011).

Pure resource-bounded autonomous constructivist systems do not exist yet in practice, but our system may be the purest to date, and almost surely is the first one that has been implemented and thus capable of providing evidence for the practicality of this otherwise theoretical stance.

3. PRINCIPLES FOR ACQUISITION OF COMMUNICATIVE SKILLS

Our approach to knowledge representation has its roots in non-axiomatic term logic and model-driven reasoning. Since knowledge in our agent is established on the basis of experience, truth is not absolute but can only be established to *a certain degree* and *within a certain time interval*. In our approach the simplest term thus encodes an observation, and is called a *fact* (or a counter-fact indicating the absence of an observation). A fact carries a *payload* (an observed event), a likelihood value in [0, 1] indicating the degree to which the fact has been ascertained and a time interval in microseconds, the period within which the fact is believed to hold (or, in the case of a 'counter-fact', the period during which the payload has not been observed). Facts have a *limited life span*, corresponding to the upper bound of their time interval. Payloads are *terms* of various types, some of which are built in the running system, the most important of these being *sensory input*, *prediction*, *goal*, *command*, *success/failure*, *internal inputs* (traces of the system's execution, enabling reflectivity), and *performance measurement*. Additionally, any of these types can be pre-defined by the system's programmer.

Except when the agent is in initial stages of bootstrapping (which should only happen once for each new environment or domain), a lot of its knowledge will be composite, that is, relationships and combinations between small "atomic" knowledge "bricks". In the case of natural language, sentences are structured out of sequences of words, with fairly complex

relationships and rules (generally called 'grammar'); words are constructed out of phonemes[2] (and letters, which have a rather complex relationship to phonemes).

To extract such knowledge from observing language-using humans in the real world the agent must have the ability to work with partially correct hypotheses about the "rules"[3] that guide the process of constructing a sentence with a particular meaning. To this end a language-learning agent would need to represent its experience as contextualized knowledge structures of some kind, with variable levels of complexity, which would allow it to change the relationships between the knowledge structures previously acquired in various ways at various levels of granularity. For instance, an incorrectly represented abstraction of how to pair nouns and verbs so that others understand what we mean might be eradicated when more examples of the various ways of its pairing are observed. The speed would be dependent on the efficiency of the agent's processes for this purpose, and this is our task here: To implement a system that can produce the necessary hypotheses for how its body and its environment works – in our case how natural language is used – and representing it in a way that allows modification to move the system towards increased accuracy. In this respect our work is compatible with e.g. that of Dominey & Boucher (2005), who demonstrated a robot learning language from limited domain and language-specific knowledge; our work goes further by proposing general principles for extracting meaning from observation, as described below (see also Nivel et al. 2014, 2013).

With an aim of generality we wanted to find a representation amenable for representing all kinds of experience; one that could be used for reasoning operations, and that would scale well by growing with vast amounts of cumulated experience – a homogenous representational scheme. Knowledge in our approach is composed of what we call *facts* (be they past, present, predicted, desired or hypothetical) and of executable code – called *models*. Models can generate new knowledge, for example predictions, assumptions, and goals, and are executable, executed at runtime by a virtual machine, the Executive. Our models are of low granularity, referred to as *peewee-size* (see Thórisson 2012, Thórisson & Nivel 2009a), each comparable in size to a SOAR's production rule (cf. Laird 2012).[4] Low granularity better supports knowledge plasticity than high granularity since modifications of small parts are less likely to have detrimental, unforeseen side-effects, and makes is easier to add/remove small parts than bigger parts, since this does less to the system. Their semantics are also simpler, and each one's effect on the entire system is easier to trace. Furthermore, peewee granularity means that higher levels of combinatorics are leveraged.

Representing time is of course necessary for producing timed behavior; for natural language time must be manipulatable at several scales, from large-scale composite operation (e.g. achieving a mission such as doing a TV interview) to intermediate-size actions (e.g. what utterance will elicit a desired answer/information from an interlocutor) to the smallest levels of individual operations (e.g. producing a prediction). This pervasion of time is a necessary

2 'Phoneme' is a construct hypothesized by humans; here it is used as shorthand for the already-categorized sounds that can be used to convey meaning in a human natural language in a modular way. Our agent is of course not bound to such human-hypothesized concepts, as it generates its own knowledge based entirely on its own experience and capabilities, provided a small seed to bootstrap the process.

3 The effective ("correct") use of natural language might be formalizable as explicit rules, but natural language is primarily a vehicle for getting things done, and as such may not be so unlike any task with complex contextual dependencies and relationships between its atomic operands.

4 While our models bear a similarity to production rules in their surface structure, having e.g. a right-hand side and a left-hand side and directly supporting reasoning, significant differences exist in other respects, including our models fusing forward and inverse control modeling supporting simultaneous and parallel forward-backward chaining, and a deep representation of time.

requirement of any system that must (a) perform in the real world and (b) model its own operation with regards to its expenditure of (temporal) resources. Considering time values as intervals allows us to encode the variable precision and accuracy needed to deal with the real world, for example, sensors do not always perform at fixed frame rates and so modeling their operation may be critical to ensure reliable operation of their controllers and models that depend on their input, and the precision for goals and predictions may vary considerably depending on both their time horizons and semantics. Last, since acquired knowledge can never be certain, one can assume that "truth" – asserting that a particular fact holds – can only be established for specific periods with varying degrees of temporal uncertainty.

4. SITUATED COGNITIVE CONTROL

Being equipped to learn natural language in situ in a social situation requires an artificial agent be endowed with many complex cognitive functions, including – among others – the ability to direct its own attention to the right things at the right time (cf. Helgason et al. 2014, Ognibene et al. 2013, Helgason et al. 2012, Demiris & Kadhouri 2006), relate spoken words and sounds to contextual actions and cues (cf. Dindo et al. 2010), and to interpret the behavior of co-actors as dependent on its underlying goals and intentions (cf. Pezzulo 2012, Dindo et al., 2011, Dennett 1987). As it bootstraps its language knowledge (from possibly meager beginnings) it needs to be capable of classifying events based on its own incomplete knowledge of the world at any point in time, in a way that it can easily update its knowledge based on gained experience.

In our approach, communicative learning, planning, and action execution are emergent processes that result from the same set of low-level processes: the execution of fine-grained programs that are automatically generated, are reusable, and are shared system-wide, collectively implementing functions that span across the entire scope of the system's operation in its environment. The most prominent program in our system is a *model*. A model is the fusion of one forward and one inverse model, according to the common terminology of control theory (cf. Wolpert & Kawato 1989), and generates both goals and predictions; some other programs monitor their success or failure and are thus able to reinforce the system's confidence about their effectiveness. Hierarchies of models represent composite knowledge and skills. The acquisition of hierarchies ameliorates attention by improving the agent's ability to anticipate, thus driving information acquisition more closely in line with contextual cues.

Knowledge in our system is *operationally constructive:* Models represent the causal relationships between observed events, be they external (sensory, i.e. reflection of the environment's states) or internal (reflection of the system's own states and operation): These are the tiny elementary executable constructs (procedural knowledge) that implement the system's ability to predict and to act in a domain. Learning a skill consists of learning models and their context and sequence of execution. A single model has three roles, (a) enabling predictions about what "may happen next", (b) suggesting specific ways to achieve a goal and, (c) making up assumptions, i.e. that while not having been observed, assert with some degree of confidence that some facts should hold given the current knowledge accumulated by the system. As practiced in control theory, hierarchies of low-level forward-inverse models (our models) constitute controllers that specify behaviors addressing environments of higher levels of complexity. Hierarchies also *compress* knowledge: The execution of models is a first class

input and captures in an abstract form (trace of computation) an otherwise complex state specification. Models thus contribute to build up increasingly abstract relationships, *grounded* in the internal system's operation ranging from surface perception/actuation complexes, to deeper (system-wise) operational semantics. At any point in time in a running system all currently relevant models are executed simultaneously in both of these roles, in parallel, in vast quantities in order to narrow down the system's options. Judgment of the system's performance relies on the continual self-assessment of its models' performance, and this controls learning: Learning happens continuously and is triggered by unexpected goal achievement or by prediction failure – learning triggering events. Upon either event, model candidates are assembled from the recorded history of salient inputs (inputs that proved of high value in the past for solving any goal pursued by the system) and fielded immediately – their relevancy at any point in time, as well as their lifetime, being sanctioned by their expected future performance. Bad models are discarded and/or replaced by better ones.

In our view, high-level processes (planning, attention, learning) influence each other reciprocally. For example, learning better models and sequences thereof improves planning; having good plans means that a system will direct its attention to more (goal-)relevant states, and this means in turn that learning is more likely to be focused on changes that impact the system's mission (e.g. correct identification of novelty), which on average increases its chances of success. These high-level processes are dynamically coupled, via the low-level processes, as they both result from the execution of the models. The system allocates computational and time resources to learning processes based on the progress of learning, i.e. the first derivative of the triggering events' rate.

At the heart of our approach is the cognitive control that results from the *continual value-driven scheduling of reasoning jobs*, the latter being small programs that perform the forward-inverse execution of models, monitor the outcomes of predictions and goals and build new models, among other tasks. High-level cognitive processes are grounded directly in the core operation of the machine, giving priority to reasoning jobs that process predictions and goals, using two complementary control schemes, top-down and bottom-up. Top-down scheduling allocates resources by estimating the global value of the jobs at hand, and this judgment results directly from the products of cognition – goals and predictions. These are relevant and accurate to various extents, depending on the quality of the knowledge accumulated so far. As the latter improves over time, goals and predictions become more relevant and accurate, allowing the system to allocate its resources with a better judgment; the most important goals and the most useful/accurate predictions have priority, the rest being saved for later processing or even discarded, to save resources. In that sense, cognition controls resource allocation. The second control scheme is bottom-up: Resource allocation controls cognition. Should resources become scarce, the scheduling process dynamically narrows the system's attention to the most important goals/predictions the system can handle, trading scope for efficiency and therefore survivability – the system will only pay attention to the most promising (value-wise) inputs and inference possibilities. If the resources become more abundant the system will start considering goals and predictions of less immediate value.

The bootstrap code – the initial *seed* for the system – contains (among other things) top-level goals (drives) and top-level models. A drive is an "innate" goal given by the programmer, whose semantics can also be those of a constraint; it is a goal whose payload is a *fact* that cannot be observed directly – think for example of the drive "keep operating successfully": the environment does not produce explicit direct evidence of its achievement, but several indicators can be combined to infer it. A top-level model is handcrafted for giving

the system a way to entail the success (or failure) from an observable (such an observable could be “your owner gives you a reward”). Due to being data-driven, drives and top-level models form together the system’s motivation, providing a top-down impetus for the system's running, while sensors provide an influx of data, driving its operation bottom-up.

A more comprehensive description of these aspects of our system, as well as others, can be found in Nivel et al. (2014, 2013), Nivel & Thórisson (2013), Nivel & Thórisson (2014), and Steunebrink et al. (2014).

5. EXPERIMENTS WITH NATURAL COMMUNICATION

The goal of the two experiments, E1 and E2, described here was to assess the ability of our first agent, S1 implemented in AERA, to learn the pragmatics, semantics, and syntax of human natural communication. We wanted an appropriately complex task that put a measure on S1's capability to autonomously disentangle a wide variety of causal relationships, sufficient to convince us about the generality of its knowledge acquisition and generalization capabilities. Human natural multimodal communication contains a wide variety of data types at two orders of magnitude of time. We defined a scenario that included considerable spatio-temporal and language behavior complexity: a dyadic mock-television interview. In the experimental setup two humans interact for some time, allowing S1 to observe their behavior and interaction in realtime; S1's task is to learn how to conduct the interaction in exactly the same way as the humans do, in either role of interviewer or interviewee. In E1 the interviewer asks the interviewee to pick up objects and move them to new locations on the table between them (Table 1), the interviewee moves the objects as requested but does not speak – a kind of *put-that-there* with learning (Bolt 1980); in E2 the interviewer asks numerous questions about the recyclability of the objects on the table between them, the interviewee giving informed answers to these (see Table 2). In E2 both interviewer and interviewee use deictics of various kinds and some forms of body language (see Table 3). A category system for non- verbal behavior was adopted from McGrew (1972), and verbal categories from Bromberg & Landré (1993). The transcribed records were then analyzed using Theme 5.0 (Magnusson 2006).

The knowledge given to S1 is represented as a small set of primitive commands for its drivers (arm joints and speech output) and categories of sensory data (speech, prosody, and joints/geometry), along with a few top-level goals such as "pleasing the interviewer" (operationally defined as the interviewer saying "thank you" or asking a new question) and "getting the interviewee to speak" (operationally defined as production of speech). The full specification of the seed for the two experiments can be found in Nivel & Thórisson (2013).

S1 observes the realtime interaction between the two humans in a virtual equivalent of a video-conference: The humans are represented as avatars in a virtual environment – to allow the interaction to proceed naturally, without any artificial protocols, each human sees the other as a realtime avatar on their screen. Their head and arm movements are tracked with motion-sensing technology (with sub-centimeter, sub-second accuracy and lag-time), their speech recorded with microphones. Signals from the motion-tracking are used to update the state of their avatars in realtime, so that everything one human does is translated virtually instantly into movements of her graphical avatar on the other's screen. Between the avatars is a desk with objects on it, visible to both participants. This is the case in both the human-human condition and the human-agent conditions (agent taking either role). In both experiments we

had S1 observe the humans until it accurately predicted all major event types observed in the dialogue (~2.5 minutes for E1 and ~20 hours for E2). We then had S1 interact with the humans for a sufficiently long period to produce videos (~10 minutes for E1, ~15 minutes for E2) that could be analyzed for t-patterns (Magnusson 2000); recordings of S1 interacting in either role with one of the humans (same as who participated in the human-human scenario) thus formed the basis for data analysis.

4.1 Experiment 1 (E1)

The objects in E1 that the interaction revolves around are: two *blue cubes*, one *red cube*, one *red sphere*, one *blue sphere*. The seed containing all initial (hand-coded) knowledge consisted of a set of primitive commands (move hand, grab, release, point at) and a set of dimensions for the input space (object type, color, actor's role, speech). The seed also includes initial knowledge that models the consequences of invoking the primitive commands: these models are for example explaining how the position of the system's hand is affected by invoking the command move hand and how a hand and an object are linked together after invoking the command grab. The natural language used in E1 consisted of a fixed set of sentence fragments (see Table 1). The seed for S1 in E1 is described in greater detail in Nivel & Thórisson (2013).

Table 1. The words and word order allowed in E1. The human participants were asked to "interact normally" to achieve their tasks (meaningless and nonsensical sentences – e.g. a sentence starting with "Take it ..." as a first sentence in an interaction, which had no prior referent for the ellipsis – did therefore not occur). We did not provide our S1 agent with any grammar or words in E1

Words	**Word Order**
verbs: put, take *nouns:* sphere, cube *adjectives:* blue, red *adverb:* there *determiners:* a, the *pronoun:* it *conjunctions:* and, ... *interjection (ack):* thank you	*Utterance:* (Part1), Part2 *Part1:* take, [a | the] noun], (conj) *Part1:* take, [it | [a | the] noun], (conj) *Part2:* put, [it | [a | the] [blue | red] noun], there, ..., thank you *(Silence of some measurable length is indicated as "..."; parenthesis means that an element is optional.)*

Results show that the performance of S1 in E1 matches the human-human scenario very closely, and S1 only needed to observe the humans for around 2.5 minutes before its performance was error-free in either role. A subsequent inspection of S1's realtime performance for 10 minutes, in realtime interaction with humans under the same operating conditions as in the human-human scenario, revealed no mistakes, restarts, or self-corrections in the interaction on behalf of S1 – it performed flawlessly and completely error-free. The system acquired and generalized interaction skills to a sufficient level to allow it to perform 100% error-free communication of the same nature and complexity as that observed in the human-human interaction.

Table 2. Basic statistics in E1. In E1 S1 only has to observe the human interaction for about 2.5 minutes before it is able to predict accurately their behavior, and subsequently assume either role without making any errors.

Observation Time	**# turns observed**	**# errors after observation period**
~2.5 minutes	~17	0

S1 learned the sequences of orders ("take a blue cube..." then waiting for the interviewee to comply before adding "...and put it there." and pointing with a finger to a location on the table), and it learned to do this with a series of different targets (e.g. a blue cube first, then a red sphere), as demonstrated by the human actors – the latter of which results from the hierarchization of control via model affordances. S1 identified the causal relationship between deictics and utterances (e.g. "there" correlated with pointing gestures) – this is an example of learned structural hierarchy – as well as ellipsis ("put it there"). The pronoun "it" was learned to identify the object that draws the most attention (in terms of learned job priority), i.e. the target of the most *valuable goals* (picking an object is a learned pre-condition on the next step, moving it to some location, to earn the reward) – this being an example of value-driven resource allocation steering cognition (and vice-versa); it matches exactly how humans used ellipsis in the observed interactions

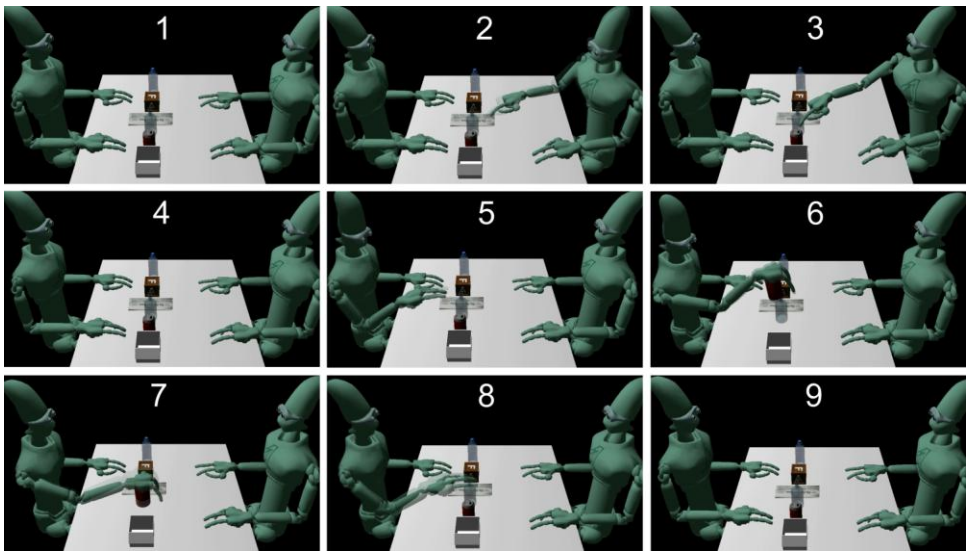

Figure 2. Example time series excerpt of interaction between human (interviewer) and S1 (interviewee) spanning seven seconds. In this interaction the human looks at an aluminum can and says "Tell me about this object," simultaneously pointing to the can with an index finger (frames 2, 3), then rests (frame 4). Approx. 600 millisecs later S1 gazes at the can and replies, grabbing the can and lifting it up, "This is an aluminum can," (frames 5, 6) puts it down again (frames 7, 8) and continues: "The main ingredient in aluminum is bauxite". This short sequence is very similar – but not identical, neither in timing nor movements of either party – to some sequences S1 had observed in the human-human interaction.

4.2 Experiment 2 (E2)

Given the success of E1, in E2 we increased the complexity of its task as follows: The scenario included all communicative behavior of E1, with a considerable increase in both spatial and language complexity. In particular, the language component in E2 included much longer and more complex sentences, and both interviewee and interviewer spoke, in full sentences and complete communicative acts. The vocabulary was 100 words; S1 was given no kind of grammar, nor a list of permissible words.[5] On the desk between the interviewer and interviewee lay a set of (virtual) objects: *aluminum can*, *glass bottle*, *plastic bottle*, *cardboard box*, *newspaper* and *painted wooden cube*. As before, the task of the participants is to talk about these objects, in particular, the interviewer's task is to ask the interviewee about the materials of which the various objects consist, and the pros, cons, cost, and methods for recycling them (Table 3). As in E1, the interviewee must understand the utterances of the interviewer to a sufficient degree to produce the desired actions, in this case long explanations about the pros and cons of recycling various kinds of materials, using deictic references, ellipsis, and standard human dialogue and turntaking skills (collaborative, non-overlapping communicative acts) (cf. Thórisson 2008, 2002). While the humans in the experiments are not trained actors and their behavior not stylized, their interaction was nevertheless correct in all major aspects – all question-answer pairs were correct and consistent. S1 thus did not have to deal with incorrect use of language, which would undoubtedly bring the observation time well above 20 hours.

Table 3. Some examples of the unscripted sentences produced in by the human participants in realtime dialogue in E2

Which releases more greenhouse gasses when produced, [an aluminum can or a glass bottle | an aluminum can or a plastic bottle | a plastic bottle or a glass bottle]?
What [else | more] can you [tell, tell me, tell us, say] about [this | that | it]?
There are many types of plastic.
Tell [me | us] about this [object | thing | one].
More energy is needed to recycle a plastic bottle than a can of aluminum.
Compared to recycled plastic, new plastic releases fifty percent more greenhouse gasses.
More energy is needed to recycle a glass bottle than a can of aluminum.
A glass bottle takes one million years to disintegrate completely in the sun.
Glass is made by melting together several minerals.
A recycled aluminum can pollutes (only) five percent of what a new [can | one] pollutes.
Recycling an aluminum can costs only five percent of a new one.
Compared to recycling, making new paper produces thirty-five percent more water pollution.
This is a cube made from unpainted wood.

5 Due to the number of commission errors in the speech recognizer, however, its output was filtered by the set of 100 words.

The results of E2 are summarized in Tables 4 and 5, and Figures 4 and 5. In E2 S1 learned everything that it observed in the human-human interactions, and can perform an equivalent interview in either role of interviewer and interviewee.[6] The full socio-communicative repertoire exemplified in E2, with additional complexity in deictic gestures and grammar, acquired autonomously by S1 after an observation period of approximately 20 hours, has been correctly learned, with no mistakes in its subsequent application, including timing of all actions (Table 4). As can be seen by comparing examples of human-S1 interaction (Figures 4, 5) with human-human interaction (Figure 3), behavioral patterns are highly significant and match closely the human-human condition, both in timing and components; the larger patterns connecting the two parties are virtually identical; disconnected smaller patterns in the S1 conditions indicate a slightly larger variation in these interactions than in the human-human condition.[7] As can also be seen clearly by simple visual inspection of the resulting videos, S1 has mastered the role of both interviewer and interviewee perfectly. T-pattern analysis revealed that the largest pattern consisting of non-overlapping hierarchical sub-patterns and found in all conditions was made up from 49 events (leaf nodes) occurring in the same order with statistically significant event timing similarity ($p < 0.005$). This largest pattern explained 77% of the total time of the interaction in the three conditions.

Table 4. Basic statistics in E2

Observation time	**# turns observed**	**# errors after observation period**
~20 hrs	~8000	0

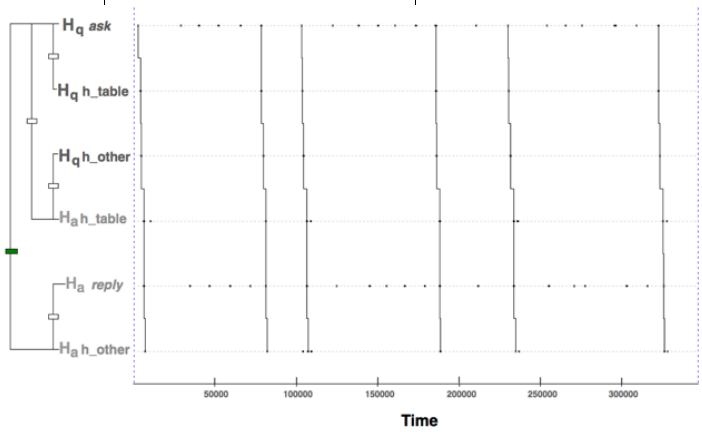

Figure 3. Example of common and statistically significant patterns seen in *human-human* condition in E2, involving question-answering, head direction, and hand activity ($p<.05$ or better). Here, the interviewer first asks a question, looks at the table, then looks back at interviewee, after which the interviewee looks at the table and begins to answer, then looks back at the interviewer. (Legend: Hq = interviewer; Ha = interviewee; ask = a question is asked; h_table = face oriented towards the objects on the table; h_other = face oriented towards interlocutor; reply = reply is produced. Timescale is in milliseconds; 5:48 mins total.)

6 Videos of the interaction can be found on www.humanobs.org and on youtube.com on channel *CADIAvideos*.

7 Note that t-pattern analysis is a mechanical mathematical procedure with no semantic labels; thus, t-patterns produced and displayed in these figures do not reflect any natural ordering such as questions preceding answers – events are simply labeled events with a beginning and end. See Thórisson et al. (2013) for an in-depth description of the t-pattern analysis used here.

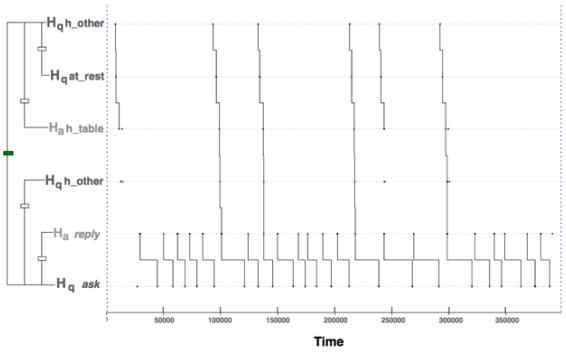

Figure 4. Example of common and statistically significant patterns seen in S1-as-interviewer condition in E2 ($p<.05$ or better). Compare this to human-human condition in Figure 3 (see text). (Legend: Hq = interviewer; Ha = interviewee; h_other = face oriented toward interlocutor; at_rest = body is at rest; h_table = face oriented towards the objects on the table; reply = answer to a question is produced; ask = a question is being asked. Timescale in milliseconds; 6:36 mins.)

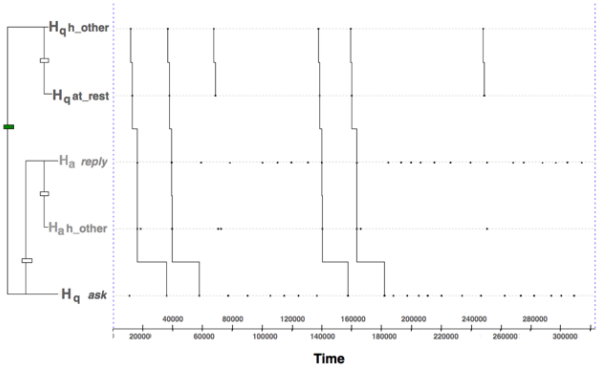

Figure 5 - Example of common and statistically significant patterns seen in human-interviews-S1 condition in E2, where questions are followed by a new question ($p<.05$ or better). Compare this to human-human condition in Figure 3 (see text). (Legend: Hq = interviewer; Ha = interviewee; h_other = facing interlocutor; at_rest = interlocutor at rest; reply = a reply to a question is produced; ask = a question is asked. Timescale is in milliseconds; 5:18 mins total.)

Table 5. Summary of results obtained in Experiment 2 (E2). S1 has learned how to conduct an interview with a human, and can perform flawlessly in either role of interviewer and interviewee after around 20 hours of observation, producing grammatically, semantically, and pragmatically correct utterances in interactions spanning minutes. Our S1 agent was not provided with any grammar or words

Category	**What Has Been Learned**	**Result**
Interview gross structure	S1 has learned how to structure dialogue in an interview, as observed in the human-human interaction. S1 has learned roles of both interviewer and interviewee from observation, having been only provided with the top-level goals for either, and can perform them both. S1 also learned to use interruption to keep the interview within the allowed time limits.	S1 can conduct dialogue with a human efficiently and effectively, as interviewer and interviewee, in a way that is virtually identical to human-human interaction. Appropriate and correct actions taken, given the behavior of either role.
Turn-taking	S1 has learned the basic skills of turn-taking from observation, as plainly obvious in the videos, and clearly demarcated in turn-taking patterns shown by t-pattern analysis. In E2 the interview includes gestures and speech for both roles. Turn-taking is slightly slower-paced than typical human-human interaction.	S1 efficiently and effectively takes turns, asking questions at the right times (as interviewer) and answers timed correctly (as interviewee). The style and action repertoire is precisely that observed in the human-human condition.
Explicit manual deictics	S1 has learned to use three kinds of deictics: pointing by finger, by palm, and picking up and putting down an object in synchrony with speech. Successful resolution of a manual deictic gesture by the interviewer allows interviewee to produce correct answer to questions, and to use it reciprocally when replying.	Both the timing and form of the gestures is appropriate for the context. Resolution of a manual deictic gesture by the interviewer allows interviewee to place objects in the right location, and to pick out a referenced object out of the five.

Ellipsis	Use of pronoun "it" and "the [*X*]" (e.g. "Take the cube" in the beginning of a new instruction) is correctly used to reference (as interviewer) / interpreted (as interviewee) an object mentioned earlier.	S1 has learned to use ellipsis in both sentence interpretation and generation. Successful resolution of ellipsis by S1 as interviewee allows it to place objects in the right location, and to pick out a referenced object out of the five.
Sentence construction	S1 constructs all sentences correctly. Correct combination of dialogue events to allow correct uses of pronoun and adverb, supporting disambiguation/ indication of what should be done.	S1 can construct sentences in either role of interviewer and interviewee, based on those observed in the human-human interaction. The sequence of words is produced using generalized models acquired autonomously from observing the human interaction.
Constructing proper answer to questions	When the interviewer asked a question, not only were the gestures and speech interpreted for the correct response, the reply constructed was appropriate to the question.	Given the numerous valid questions that can be asked in E2, S1 replies with an appropriate and correct utterance.

6. CONCLUSION

We have demonstrated an implemented architecture that can learn autonomously many things in parallel, at multiple time scales. The results show that the AERA-based S1 agent can learn complex multi-dimensional tasks from observation from only a small ontology, a few drives (high-level goals), and a few initial domain models to support autonomous bootstrapping on a complex task. Human dialogue is an excellent example of the kinds of complex tasks current systems are incapable of handling autonomously, and to our knowledge no prior architecture has demonstrated comparable results (cf. Franklin et al. 2013, Laird 2012, Wang 2011). The fact that no difference of any importance can be seen in the performance between S1 and the humans in simulated face-to-face interview is an indication that the resulting architecture holds significant potential for further advances, and that our methodology (Nivel et al. 2013, Thórisson 2012) is a way for escaping the constraints of current computer science and engineering software methodologies when aiming for artificial general intelligence and increased systems autonomy. However, in its current incarnation AERA is entirely dependent on observation, as learning is exclusively triggered by unexpected goal achievement, or a prediction that turns out to be wrong – i.e. by surprise. This limits the acquisition of knowledge to phenomena that are directly observable – hidden causation is difficult for the current system to figure out, as are other kinds of inexplicit relations (similarity, equivalence, etc.). Elsewhere we have argued that curiosity results from the need to overcome the limitations imposed by the scarcity of inputs (Steunebrink et al. 2013); we plan to expand the types of programs to implement a richer set of inferences from which curious behaviors can be devised and planned, whenever the system has resources to spare. One of the main directions of our planned near-future work is set toward building more prototypes to assess the generality and scalability of our system.

ACKNOWLEDGEMENT

This work was supported by the European Project HUMANOBS – Humanoids that Learn Socio- Communicative Skills By Observation (FP7 STREP – Cognitive Robotics, Grant number 231453), by Nascence (FP7-ICT-317662), SNF grant #200020-138219, and by research grants managed by Rannís, Iceland. We are grateful to H. Th. Thorisson and G. S. Valgardsson for the design of the S1 avatar and Th. Bryndis Thorisdottir for researching the E2 interview content.

REFERENCES

Bolt, R. A. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. Computer Graphics, Vol. 14 No. 3, 262-70.

Demiris Y., Khadhouri B. 2006. Hierarchical attentive multiple models for execution and recognition of actions. Robotics and autonomous systems. Vol. 54, pp. 361-369.

Dennett, D. 1987. *The intentional stance*. MIT Press, Cambridge, Massachusetts.

Dindo, H.; Zambuto, D. & Pezzulo, G. 2011. Motor simulation via coupled internal models using sequential Monte Carlo. *Proceedings of IJCAI 2011*, pp. 2113-2119.

Dindo, H. and D. Zambuto. 2010. A Probabilistic Approach to Learning a Visually Grounded Language Model through Human-Robot Interaction. *2010 IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS)*, Taipei, Taiwan, October 18-22, 2010.

Dominey, P. F. & J-D. Boucher 2005. Developmental Stages of Perception and Language Acquisition in a Perceptually Grounded Robot. *Cogn. Syst. Res.*, Vol. 6 No. 3, pp. 243-259.

Franklin, S., T. Madl, D'Mello, K. Sidney & J. Snaider 2013. LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning. *Transactions on Autonomous Mental Development*.

Goertzel, B., C. Pennachin, S. Araujo, F. Silva 2013. A General Intelligence-Oriented Architecture for Embodied Natural Language Processing.

Helgason, H. P., Thórisson, K. R., Garrett, D. & Nivel, E. 2014. Towards a General Attention Mechanism for Embedded Intelligent Systems. International Journal of Computer Science and Artificial Intelligence 03/2014; Vol. 4, No. 1, pp. 1-7.

Helgason, H. P., Nivel, E. & Thórisson, K. R. 2012. On attention mechanisms for AGI: A design proposal. *Proceedings of Artificial General Intelligence conference* (AGI'12), pp. 89-98.

Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press, Cambridge, MA.

McGrew, W. C. (1972). *An Ethological Study of Children's Behaviour*. New York: Lawrence Erlbaum Associates.

Magnusson, M. S. 2000. Discovering hidden time Patterns in Behavior: T-patterns and their detection. *Behavior Research Methods, Instruments & Computers*, Vol. 32, pp. 93-110.

Nivel, E. & Thórisson, K. R. 2009. Self-Programming: Operationalizing Autonomy. *Proceedings of the Second Conference on Artificial General Intelligence*. 2009.

Nivel, E. & K. R. Thórisson 2013. Seed Specification for AERA S1 in Experiments 1 & 2. Reykjavik University School of Computer Science Technical Report, RUTR-SCS13005.

Nivel, E., K. R. Thórisson, B. R. Steunebrink, H. Dindo, G. Pezzulo, M. Rodriguez, C. Hernandez, D. Ognibene, J. Schmidhuber, R. Sanz, H. P. Helgason, A. Chella & G. K. Jonsson 2013. Bounded Recursive Self-Improvement. RU-SCS130006 Technical Report (ArXiV: 1312.6764).

Nivel, E., K. R. Thórisson, B. R. Steunebrink.,H. Dindo, G. Peluzo, M. Rodriguez, C. Hernandez, D. Ognibene, J. Schmidhuber, R. Sanz, H. P. Helgason & A. Chella 2014. Bounded Seed-AGI. *Proceedings of Artificial General Intelligence* (AGI-14), pp. 85-96, Quebec, Canada.

Nivel, E. & Thórisson, K. R. 2013. Towards a programming paradigm for control systems with high levels of existential autonomy. *Proceedings of Artificial General Intelligence* (AGI-13), pp. 78-87, Beijing, China.

Ognibene, D. & Demeris, Y. 2013. Active event recognition. *Proceeding, 23rd International Joint Conference of Artificial Intelligence* (IJCAI13), pp. 2495-2501.

Ognibene, D., Chinellato, E., Sarabia, M. & Demiris, Y. 2013. Contextual action recognition and target localisation with active allocation of attention on a humanoid robot. *Bioinspiration & Biomimetics*, Vol. 8 No. 3, 035002. doi:10.1088/1748-3182/8/3/035002.

Petit, M., Lallee, S., Boucher, J.-D., Pointeau, G., Cheminade, P., Ognibene, D, Chinellato, E., Pattacini, U., Gori, I., Martinez-Hernandez, U., Barron-Gonzalez, H., Inderbitzin, M., Luvizotto, A., Vouloutsi, V., Demiris, Y., Metta, G. & Dominey, P.F. 2012. The Coordinating Role of Language in Real-Time Multimodal Learning of Cooperative Tasks. IEEE Transactions on Autonomous Mental Development, Vol. 5, No. 1, pp. 3-17.

Pezzulo, G. 2012. The Interaction Engine: A common pragmatic competence across linguistic and non-linguistic interactions. *IEEE Transactions on Autonomous Mental Development*, Vol. 4, pp. 105-123.

Piaget, J. 1950. *The Psychology of Intelligence*. Routledge & Kegan Paul, London, England.

Steunebrink, B. R., J. Koutnik, K. R. Thórisson, E. Nivel & J. Schmidhuber 2013. Resource-Bounded Machines are Motivated to be Efficient, Effective, and Curious. In K-U Kühnberger, S. Rudolph and P. Wang (eds.), *Proceedings of the Sixth Conference on Artificial General Intelligence (AGI-13*), pp. 119-129, Beijing, China.

Thórisson, K. R. 2002. Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. B. Granström, D. House, I. Karlsson (Eds.), *Multimodality in Language and Speech Systems*, pp. 173-207. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Thórisson, K. R. 2008. Modeling Multimodal Communication as a Complex System. I. Wachsmuth, M. Lenzen, G. Knoblich (eds.), *Springer Lecture Series in Computer Science: Modeling Communication with Robots and Virtual Humans*, pp. 143-168. Springer, New York.

Thórisson, K. R. & Nivel, E. 2009a. Achieving Artificial General Intelligence Through Peewee Granularity. *Proceedings of the Second Conference on Artificial General Intelligence*, 222-223, Arlington, VA, USA, March 6-9

Thórisson, K. R. & Nivel, E. 2009b. Holistic Intelligence: Transversal Skills and Current Methodologies. *Proceedings of the Second Conference on Artificial General Intelligence*, 220-221, Arlington, VA, USA, March 6-9

Thórisson, K. R. 2012. A New Constructivist AI: From Manual Construction to Self-Constructive Systems. P. Wang & B. Goertzel (eds.), *Theoretical Foundations of Artificial General Intelligence*, Vol. 4, pp. 145-171. Springer, New York.

Thórisson, K. R. 2013. Reductio ad Absurdum: On Oversimplification in Computer Science and its Pernicious Effect on Artificial Intelligence Research. In A. H. M. Abdel-Fattah &K.-U. Kühnberger (eds.), *Proceedings of the Workshop Formalizing Mechanisms for Artificial General Intelligence and Cognition* (Formal MAGiC), Beijing, China, July 31st, 31-35. Institute of Cognitive Science, Osnabrück.

Thórisson, K. R., Nivel, E., Pezzulo, G., Ognibene, D., Magnusson, M. S. & Jonsson, G. K. 2013. Evaluating AGI-aspiring systems via human-robot interaction using t-patterns. Reykjavik University School of Computer Science Technical Report, RUTR-CS13004.

Wang, P. (2004). Toward a unified artificial intelligence. *Papers from the 2004 AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Research and Systems*, pp. 83–90.

Wang, P. 2006. *Rigid Flexibility: The Logic of Intelligence*. Springer, Dordrecht.

Wang, P. 2011. The assumptions on knowledge and resources in models of rationality. *International Journal of Machine Consciousness*, Vol. 3, No. 1, pp. 193-218.

Wolpert, D. M. & Kawato, M. 1998. Mutliple paired forward and inverse models for motor control. Neural Networks, Vol. 11, pp. 1317–1329.