About Understanding

Kristinn R. Thórisson[1,2(✉)], David Kremelberg[2], Bas R. Steunebrink[3], and Eric Nivel[2]

[1] Center for Analysis and Design of Intelligent Agents, Reykjavik University, Reykjavik, Iceland
thorisson@ru.is
[2] Icelandic Institute for Intelligent Machines, Reykjavik, Iceland
[3] The Swiss AI Lab IDSIA, USI and SUPSI, Manno, Switzerland

Abstract. The concept of *understanding* is commonly used in everyday communications, and seems to lie at the heart of human intelligence. However, no concrete theory of understanding has been fielded as of yet in artificial intelligence (AI), and references on this subject are far from abundant in the research literature. We contend that the ability of an artificial system to autonomously deepen its understanding of phenomena in its surroundings must be part of any system design targeting general intelligence. We present a theory of *pragmatic understanding*, discuss its implications for architectural design and analyze the behavior of an intelligent agent implementing the theory. Our agent learns to understand *how to* perform multimodal dialogue with humans through observation, becoming capable of constructing sentences with complex grammar, generating proper question-answer patterns, correctly resolving and generating anaphora with coordinated deictic gestures, producing efficient turntaking, and following the structure of interviews, without any information on this being provided up front.

1 Introduction

A rudimentary investigation into the use of the term "understanding" in the field of artificial intelligence (AI) reveals that occurrences are few and far between. When it does appear it is primarily in the context of natural language ("language understanding"), where parsing and manipulation of linguistic tokens (read: good old-fashioned AI) takes the front seat. A distant second is its coupling with the words "scene" and "image" in computer vision research (scene understanding, image understanding), with an identical emphasis on parsing: Rather than talking about the phenomenon of understanding proper, understanding is equated with syntactic manipulation, which, as everyone who has studied philosophy knows, is not the same thing (cf. [18]).

A coherent conceptualization of understanding is of importance to the field of AGI for several reasons. First, if the concept of understanding is left undefined it cannot, as a phenomenon, be effectively investigated; second, without a good definition of understanding it may be difficult to compare different systems with respect to their level of understanding, and similarly, to compare the same

B. Steunebrink et al. (Eds.): AGI 2016, LNAI 9782, pp. 106–117, 2016.
DOI: 10.1007/978-3-319-41649-6_11

system or different systems with respect to their levels of understanding regarding different areas of expertise or performance; and third, a coherent account of understanding is needed such that system builders can create new systems, improve current systems, and train systems where understanding is a specific goal. A formalized account of understanding would seem crucial to the continued and successful progress of the field of AGI.

The apparent indifference of AI researchers to the phenomenon of understanding is curious considering the available evidence about its role in human intelligence. If "understanding" is simply a descriptive term used to classify the effectiveness of a given behavior for a particular goal, after it has been observed – behavior referring here to perception, thinking, and action control – then perhaps it could be said that intelligence and understanding are synonyms, and ignoring the concept altogether is justified. If, however, understanding is a unique ingredient or property of natural thinking systems which affects their abilities and intelligence – and especially: their potential for growing their own knowledge – then we would be well-served by studying understanding as a phenomenon. We argue for the latter view and outline here a *pragmatic theory of understanding* rooted in an analysis of how predictive controllers compute meaning. First we look at some of the relevant background work from philosophy and AI, then we present our theory of pragmatic understanding and meaning, and then give an overview of the results of a prototype system whose knowledge acquisition and application was constructed according to the theory. The results represent strong evidence for the potential of the theory to elucidate the relationship between meaning, understanding, prediction, and explanation, in a manner relevant to artificial general intelligence.

2 Related Work

An important question that has been discussed, mostly in the philosophical literature, is the extent to which machines could be given understanding, if at all. Sloman has stated that the question of whether machines can "really" understand is more of a minor question of definition than anything else [18], arguing that the appropriate answer to the question "Can you understand?" is not binary and can take the form of infinite features and gradations. It seems a latent view of many that once a machine can do some human task, that task is no longer deemed as requiring "intelligence," and by extension, requires no "real understanding". This view might explain why Searle's Chinese Room argument still has appeal, in spite of the numerous publications that have long since refuted it by illustrating its numerous fallacies [2,3,16]. Convincing arguments for the impossibility of machines to understand remain scarce.

Some research has argued for the importance of understanding in cognition, citing it as distinct from knowledge (cf. [6]), claiming that acquisition (deepening) of understanding constitutes a more accurate reflection of the world than knowledge acquisition [7,8], and is thus a greater intellectual achievement. Others have taken the exact opposing view (cf. [10]). Without proper and reasonably

specific definitions of these terms and their context, as these accounts tend to be, they can be somewhat incoherent, too heavily steered by the many senses in which the term might be used colloquially. As a result many seemingly irreconcilable polarities and contradictions are uncovered [9] (for instance, pitting the internal organization of phenomena against its relation to various other phenomena as some sort of contradiction). As we shall attempt to demonstrate below, such inconsistencies may be reconciled with proper definitions and the right unifying approach.

While understanding as a phenomenon has received more attention in the philosophical than the AI literature [7,8], even there it has nevertheless been claimed to have "virtually escaped investigation in English-speaking philosophy" ([5]: 307); this dearth of interest in the subject is evidenced not only there but also in the fields of AI and cognitive science.[1] A few books have been published with the word pair "understanding understanding" in the title [4,15]. Interesting as they may be, one of these contains selected writings by cybernetics pioneer Heinz von Forrester, which, in spite of its promising title, is not about understanding at all (as evidenced by the word "understanding" not appearing the index); the other gives a cursory (albeit a decent) summary of the subject in the context of epistemological philosophy.

In the context of the work presented here, few authors if any have addressed the more relevant question of what kinds of architectures could *deepen their understanding* automatically, as this would seem of key importance for an AGI system for growing its knowledge. Here we attempt a unification of several prior ideas, through the concepts of prediction, granular model generation and evaluation, and knowledge acquisition through experience [19]. While the literature has presented a multitude of ways to look at and define understanding, and virtually all of the concepts we talk about have appeared in the AI literature in one form or another, we are not aware of any that propose the kind of unification presented here.

3 Towards a Theory of Pragmatic Understanding

Our concern here is with an agent's understanding of phenomena of interest that allows it to act intelligently towards it, in a practical and goal-directed way. We refer to our theory of understanding as *pragmatic*, as we are concerned with the *usefulness* that levels of understanding may achieve in guiding behavior.

Phenomenon. A phenomenon Φ (process, state of affairs, occurrence) — where W is the world and $\Phi \subset W$ — is made up of a set of elements[2] $\{\varphi_1 \ldots \varphi_n \in \Phi\}$

[1] Exceptions do exist of course (cf. [1]), but not in the obvious areas such as language-, image- and scene-understanding, where the word makes a mere superfluous appearance.

[2] By "elements" and "sub-parts" we mean any sub-division of Φ, including substructures, component processes, whole-part relations, causal relations, etc.

of various kinds including relations $\Re_\Phi$ (causal, mereological, etc.) that couple elements of Φ with each other, and with those of other phenomena.

Phenomenon and Context. The relations $\Re_\Phi \subseteq 2^W \times 2^W$ that extend to other phenomena identify the phenomenon's *context*. We partition $\Re_\Phi$ in *inward facing* relations $\Re^{in}_\Phi = \Re_\Phi \cap (2^\Phi \times 2^\Phi)$ and *outward facing* relations $\Re^{out}_\Phi = \Re_\Phi \setminus \Re^{in}_\Phi$. An agent whose models are only accurate for $\Re^{in}_\Phi$ understands Φ but not Φ's relation to other phenomena; an agent whose models are only accurate for $\Re^{out}_\Phi$ understands Φ's relation to other phenomena but will have limited or no understanding of Φ's internals.

Models. M_Φ is a set containing models of a phenomenon Φ $\{m_1 \dots m_n \in M_\Phi\}$ – information structures that can be used to (a) *explain* Φ, (b) *predict* Φ, (c) produce effective plans for achieving goals G with respect to Φ, and (d) (re)create Φ.

For any set of models M and a phenomenon Φ, the closer the information structures $m_i \in M$ represent elements (sub-parts) $\varphi \in \Phi$, at any level of detail, including their couplings $\Re_\Phi$, the greater the *accuracy* of M with respect to Φ.

Insofar as an agent A's knowledge consists of models M, we can define *understanding* in the following way:

Understanding. An agent A's *understanding* of phenomenon Φ depends on the accuracy of M with respect to Φ, M_Φ. Understanding is a (multidimensional) *gradient* from low to high levels, determined by the quality (correctness) of representation of two main factors in M_Φ:

U1: The *completeness* of the set of elements $\varphi \in \Phi$ represented by M_Φ.
U2: The *accuracy* of the relevant elements φ represented by M_Φ.

Testing for Understanding. This approach does not necessitate or force any particular way to test for understanding, shifting that challenge rather to whichever methods prove the best for exposing the above two factors. To test for evidence of understanding a phenomenon Φ we may probe (at least) four capabilities of the understander:

1. To *predict* Φ.
2. To *achieve goals* with respect to Φ.
3. To *explain* Φ.
4. To *(re)create* Φ.

All can be seen to have a range $[0, 1]$ where 0 is no ability and 1 is perfection, as a function on **U1** and **U2** above. For a thorough evaluation of understanding all four should be applied.

Prediction is the crudest form of evidence for understanding. Some prediction can be done based on correlations, as prediction does not require representation of the direction of causation yet captures co-occurrence of events. Prediction of a particular turn of events requires (a) setting up initial variables correctly,

and (b) simulating the implications of (computing deductions from) this initial setup.

Goal Achievement Correlation is not sufficient, however, to inform how one achieves goals with respect to some phenomenon Φ. For this one needs causal relations. Achieving goals means that some variables in Φ can be manipulated directly (or indirectly via intermediate variables). Unless the intelligent agent is omnipotent and omniscient, to achieve goals with respect to a phenomenon Φ may require a bit more than an understanding of Φ: it requires understanding of how a certain subset of Φ relates to some variables that are *under an agent's control*. In short, the agent needs models for interaction with the world. For a robotic agent driving a regular automobile, to take one example, the agent must possess models of its own sensors and manipulators and how these relate to the automobile's controls (steering wheel, brakes, accelerator, etc.). Such interfaces tend to be rather task-specific, however, and are thus undesirable as a required part of an evaluation scheme for understanding. Instead, we call for an ability to *produce effective plans* for achieving goals with respect to Φ. An effective plan is one that can be proven useful, efficient, effective, and correct, through implementation.[3]

Explanation is an even stronger requirement for demonstrating understanding. Correlation does not imply causation, which means that one may have a predictive model of a phenomenon that nevertheless does not represent correctly its parts and their relations (to each other and parts of other phenomena); goals may in some cases be achieved through "hacks" and "back doors", without a proper causal model behind it. This is why scientific models and theories must be both predictive *and* explanatory – together constituting a litmus test for complete and accurate capturing of causal relations.

(Re)creating a phenomenon is perhaps the strongest kind of evidence for understanding. It is also a pre-requisite for the ability for correctly building new knowledge that relies on it, which in turn is the key to growing one's understanding of the world. By "creating" we mean, as in the case of noted physicist Richard Feynman,[4] the ability to produce a model of the phenomenon in sufficient detail to replicate its necessary and sufficient features. Requiring understanders to produce models exposes the completeness of their understanding.

It is important to emphasize here that understanding, in this formulation, is not reductionist: Neither does it equate the ability to understand with the ability to behave in certain ways toward a phenomenon (e.g. achieve goals), nor the ability to predict it, nor the ability to explain it, nor the ability to (re)create it. While any of these may be used to assess a system's understanding of a

[3] Producing plans, while not being as specific as requiring intimate familiarity with some I/O devices to every Φ, requires nevertheless knowledge of some language for producing said plans, but it is somewhat more general and thus probably a better choice.

[4] Feynman, notorious for his capacity to understand even the most complicated phenomena in his field, left a note on his blackboard when he died: "What I cannot create, I do not understand." (http://archives-dc.library.caltech.edu/islandora/object/ct1:483 - accessed Apr 2, 2016).

phenomenon, in our theory *all are really required* (to some minimum extent) to (properly) assess a system's understanding. Any assessment method that does not include these four in some form runs the risk of concluding understanding where there is none (and the converse).

4 Meaning

We can now move to a close cousin of understanding – *meaning*. Meaning does not exist in a vacuum: A causal event x acquires meaning for some agent A when x has potential to influence something of relevance to one or more of the agent's goals G. Given e.g. an event x with potential relevance to agent A, the agent may *compute* some meaning of x with respect to (any or all of) its relevant goals, given a particular situation S_t (a substate of a world W defined by a set of variables, $S \subset W$). This computation relies on deduction, among other processes.

To illustrate we can use two example events, rocks rolling down a hill and a computer deriving square roots. Do rocks rolling down a mountainside contain any meaning? When a computer is given the number 4 and outputs 2, does this output have any meaning? "Surely", you might be inclined to say, "math is meaningful in its regularity". But then what is the difference between computation and rolling rocks? At the atomic level are forces at play (gravity and electricity, respectively) working according to predetermined rules. To answer either question we must ask "meaning to *whom*?" Both are physical events, and without a biological being that can interpret them in some relevant context, neither has any meaning.

As we can see from this example, the agent's situation must also be included, because some event x may mean one thing in situation S_1 and another in situation S_2. If I hear an announcement that the gate to the flight to my vacation destination has closed, this will mean something very different depending on which side of the gate I am on at that point in time; in one case I may start crying and the other not. And if I have a drink in either contingency it will likely be for very different reasons. This example makes another aspect of meaning clear: Meaning is time-dependent.

This means that without temporally demarcated goals there can be no meaning, because the meaning of e.g. an event can only exist with respect to a particular goal (held by an agent) that is relevant to the agent. A stone rolling down a hill has no meaning – it is simply a meaningless process. When we know the stone weighs over two tons and it's heading your way do we derive some meaning from its existence.

In this formulation the meaning of a particular datum,[5] e.g. the closing of the gate, consists of the *implications* $I_{d(t1)}$ of that particular datum d presented at time $t1$; $d(t1)$ *implies* some set of things for a particular agent A in particular

[5] A datum d_t can be an event, an utterance, the perception of a particular object, a particular deduction or set of deductions, etc. occurring at time t – in short, anything that can be perceived by the agent's sensors and represented by its mind.

circumstances S_{t1} with regard to particular active goals G (an active goal at time t is a goal that the agent is actively trying to achieve at time t).[6] Any potential implication may be computed through the proper processes, including implications that might be *relevant* to the agent's active goals G in situation S at time t_1. To be as useful as possible to the agent, the implications that are *most temporally relevant* to the agent's goals, whether a hindrance or help, should get computed as soon as possible after the datum presents itself.

Implications are computed through temporally-grounded deduction, from a set of premises, to derive any potential implications (they are *potential* implications because they are typically produced based on premises and initial conditions whose specification may not be fully informed) given by the new datum. For instance, if I missed my airplane and the next airplane leaves in a week, I may have shortened my vacation by 50 %. In this case knowing this 400 ms sooner or 4 s later will obviously not make a big difference – either way I will be steaming angry or hugely disappointed, as the meaning is extracted and the most relevant implications for my goal of taking a 2-week vacation dawns on me.

Implications. Starting from an initial state $S_t \subset W$ of a dynamic task-environment (consisting of a series of such states $\{S_t \ldots S_{t+\delta}\}$), the *Implications* of a datum d_t are the computed deductions D that may be relevant to a particular set of goals G of a particular agent A with particular knowledge K in situation $S_{t+i} \subset W$, represented

$$Impl(d_t, A(G)_{t+x}) = D(d_t, S_i, (K_A, G_A, S_A)_{t+y})$$

($t+x$ and $t+y$ means these can refer to different points in time). While for any period of time at least some implication can be deduced from a particular set of information, whether the implications are relevant to an agent cannot be known before the deductions have been made.

Most of the time a complex environment such as the physical world will present, for any time period, a vastly greater amount of information than what any agent can perceive and process for that period, i.e. the computational resources of most (interesting) agents will be vastly less than those needed to process all available information, for any time period. In the vast majority of cases such a complex environment can be the source of an infinite string of deductions stretching into the far future; for any time interval a real agent in a real environment will thus be faced with capping deductions in both breadth (sources of deductions) and depth (time and detail).

For an agent, finding the meaning of a situation requires identifying which of the possible deductions are relevant to the agent's goals in that situation at that time.

[6] Unless otherwise specified the term "goal" may be read to mean "all active goals", as typically this is *a set* of goals; even if a single identifiable top-level goal can be found, there will always be (obvious and non-obvious) sub-goals that must be taken into account. We thus use "goal" and "goals" indiscriminately.

Meaning. The *meaning* of a datum d_t for an agent $A(K, G, S)_t$ is captured by the set of *relevant implications* I_r of d_t for A with a set of goals G and knowledge K in situation S at time t;

$$Meaning(d_t, A(G)) = Impl_r(d_{t+x}, \{K_A, G_A, S_A\}_{t+y}).$$

Typically there is never only a single meaning to anything (so we use singular and plural interchangeably), since any datum has a large set of potential implications for any large or complex phenomenon. What is *relevant* at any point in time depends on the particular outcome of the predictions, in light of the system's active goals. Since these predictions cannot be guaranteed to be perfect, the meaning of anything and everything will always be somewhat in flux and open to further interpretation. Computations may produce differences in meaning based on slight variations of the initial conditions.

The quality of predictions produced via deductions from a set of premises depends in large part on the accuracy of the models used for it. Models must be freely composable and de-composable, in light of their usage, to realize their full potential for predicting, achieving goals, and explaining. From Ashby's Requisite Variety theorem [17] we know that model "resolution" (i.e. their granularity) needs to be at least as detailed as the finest discernible, relevant details of the phenomenon modeled. For any reasonably complex phenomenon we will therefore have a large set of models M.[7]

5 A System that Acquires Understanding and Meaning

We have designed and implemented an architecture that implements the pragmatic theory of understanding outlined above. This system, called AERA [12,13], contains numerous features that must be explained to provide a coherent account of its operation, which is well beyond the scope of this paper (we refer the interested reader to our most thorough overview of this work in [11]). Rather, this section serves (a) to show that our approach to understanding and meaning has produced an implemented, working system, (b) to show that this system demonstrates highly novel properties not seen before in any other system, and perhaps most importantly, (c) to show one way the above theory can be mapped to a concrete implementation.

Based on a new constructivist methodology [20], an AERA agent can learn complex tasks by observation, starting from only a tiny seed. Learning in AERA is life-long, continuous, and incremental, and consists of building models based on observed phenomena. For any situation $S_i(t) \subset W$ the system finds itself

[7] Another determinant of the quality of predictions is the observability of variables and the accuracy of reading their values. For any triplet $\{A, G, S\}$, to produce predictions requires fixing the values of numerous variables $v \in V \subset S$ whose values may not be immediately accessible (and thus guessed or retrieved from the agent's prior experience), or whose values may not be perfectly observable (cf. "Does that display show 880 or 830?").

in, a set of observed variables $V_i \subseteq S_i$ results in a large set of new models M_i being generated, each relating two observed data $v_i, v_j \in V$ in a directed causal relationship $\Re_i : v_i \rightarrow v_j$, meaning that v_i is a cause of v_j. As experience accumulates, models of groupings of such relationship pairs emerge, representing hypotheses about the interactions between the many observed sub-phenomena, at several spatio-temporal levels. At runtime an AERA agent executes the subset of these models deemed most relevant to the situation; predictions are produced from the present state using these models for deduction, in a forward-chaining mode; abduction — backward-chaining the models' causal relationships — produces plans for how to achieve goals (i.e. partial world states not observed at present).

In the experimental data referenced below, AERA agent S1's phenomenon Φ to be understood is a TV-style interview. This Φ's elements are known to be e.g. deictic references (pointing at, nodding towards, looking at, etc.), sentence morphology (word sequences), question-answer pairs, etc. S1 starts with a tiny seed where its most primitive sensation types are specified, allowing it to ground its experience and bootstrap its incremental learning of how to properly do multimodal interaction. The seed also contains the top-level goals (1 for the interviewer, 4 for the interviewee). S1 observes two humans interact for 20 h, after which its performance is recorded for analysis, producing over 20 min of interactions with humans. It is important to note that no information whatsoever was provided in the seed on any of the phenomena learned – these emerge through a process whereby the system tries to match its models to the observed phenomena in a way that can predict, explain, and achieve goals with respect to them, as per our pragmatic theory of understanding detailed above.

Explanation. By design the system's knowledge representation is self-disclosing: The total collection of models at any point in time represents the system's ability to explain the phenomena it has had experience with, from its best effort, by attempting to represent directly the elements of the phenomenon (observable variables) and their relationships ($\Re^{in}_{\Phi}$). This is very different from e.g. artificial neural nets, whose knowledge representation cannot be symbolically mapped to the domain the knowledge references.

Prediction. Models and model hierarchies are used to predict the evolution of the situation, at any moment, δ microseconds into the future $S_i(t_{now} + \delta)$. Models get a score according to how closely the observed future compares to their predictions.

Goal Achievement. The same models used to produce predictions also inform the system what it is capable of, via backward chaining: Any (good) model chain of arbitrary length whose end point is a goal to be achieved and whose starting point is the present state tells the system what chain of events may be taken to get from the present state to the goal, and as long as the chain includes models referencing variables that the system can affect, the system can create a plan for achieving the goal. In such chains the agent's atomic operational capabilities are

represented in models, and their execution is handled via dedicated actuators on the agent's embodiment.

Implications and Meaning. Having acquired a set of models, when an AERA agent observes a datum $d_i(t)$ the best models in which this datum appears produces predictions (arity depending on available resources); those that relate in some way to the agent's instantiated goals at that point in time are considered relevant implications of d_i, and may affect the agent's subsequent overt actions in the task-environment. Meaning is thus generated continuously, with the predictions most relevant to the agent at each point in time enabling the agent to steer its behavior accordingly, by changing its plans, creating new plans, backtracking, and abandoning or generating new subgoals.

Results: Autonomously Acquired Understanding. In two experiments S1 has demonstrated *autonomous acquisition of a pragmatic understanding* of (1) three types of linguistic anaphora (resolving referents of "it", "that" and "this"), (2) four types of co-verbal deictic gestures (pointing with index finger, gazing at objects, palm-up hand gesture, reference via touching/holding objects), (3) how to structure turn-taking, (4) how to generate appropriate utterances for particular referents (correct answers to questions – whether containing anaphora, co-verbal gestures or not), (5) how to keep an interaction within given time limits, and (6) how to generate syntactically correct utterances. With respect to item 6, examples of utterances produced by S1 include "Which releases more greenhouse gases when produced, a plastic bottle or a glass bottle?" and "Compared to recycling, making new paper results in seventy five percent more air pollution." As evidence of the accuracy and completeness of S1's understanding, for the total of 73 utterances produced by S1 in the experimental data, only four (minor) grammatical errors were found (Nivel et al. [12] provides details S1's natural language learning.). These were in fact the *only* errors found — no errors could be discerned in the data for any of the other acquired skills (1–5).

This evidence suggests that with respect to the sub-phenomena listed above, all the above elements of Φ have been modeled correctly, achieving a high score on U1 and U2 in Sect. 3. The first two points of evaluation in Sect. 3 are thus clearly demonstrated: The system can use its acquired understanding to achieve goals in the dialogue, using both prediction (to synchronize behavior with the world) and abduction (to construct plans). We consider items 3 and 4, explanation and (re)creation, to partially demonstrated: S1's self-disclosing knowledge representation directly captures the structure of the phenomena by encoding the (causal) relationships between observed variables, and allows S1 to act correctly across the full range of priorly observed instances of the phenomena. More thorough evaluation is needed on these last two points, including pushing the limits of S1's understanding.

6 Conclusions

We have outlined a theory of pragmatic understanding and meaning. The implemented system incorporating its principles lends validity to the approach, and for

the more general issue that endowing agents with capabilities for autonomously acquiring a pragmatic understanding of a complex phenomenon may be an important endeavor. The implemented system has demonstrated an ability to acquire complex sentence grammar from observation, contextual interpretation of multimodal communicative acts, acquiring an understanding of a task-environment and computing in real-time the meaning of events, and using this to successfully achieve dialogue goals in realtime interaction with humans [11,12]. In this the system demonstrates what Pattee calls *semantic closure* [14]. Needless to say, the issue of understanding is a large one, and a multitude of issues have been raised here that remain unformulated, let alone unanswered, such as susceptibility to noise and scaling. The positive results from our experiments thus far provide good reason for optimism on the future prospects of this line of research.

Acknowledgments. We would like to thank our HUMANOBS collaborators' valuable contributions to the AERA system. This work was sponsored in part by the School of Computer Science at Reykjavik University, by a European Project HUMANOBS (FP7 STREP #231453), by a Centers of Excellence Grant from the Science & Technology Policy Council of Iceland, and by a grant from the Future of Life Institute.

References

1. Baum, E.: Project to build programs that understand. In: Proceedings of the Second Conference on Artificial General Intelligence, pp. 1–6 (2009)
2. Chalmers, D.J.: Subsymbolic computation and the chinese room. In: Dinsmore, J. (ed.) The Symbolic and Connectionist Paradigms: Closing the Gap. Lawrence Erlbaum, Hillsdale (1992)
3. Fisher, J.A.: The wrong stuff: chinese rooms and the nature of understanding. Philos. Inves. **11**(4), 279–299 (1988)
4. von Forrester, H.: Understanding Understanding: Essays on Cybernetics and Cognition. Springer, New York (2003)
5. Franklin, R.L.: On understanding. Philos. Phenomenological Res. **43**(3), 307–328 (1983)
6. de Gelder, B.: I know what you mean, but if only i understood you. . . In: Parret, H., Bouveresse, J. (eds.) Meaning and Understanding, pp. 44–61. de Gruyter, Berlin (1981)
7. Grimm, S.R.: The value of understanding. Philos. Compass **7**(2), 279–299 (1988)
8. Grimm, S.R.: Understanding as knowledge of causes. In: Fairweather, A. (ed.) Virtue Epistemology Naturalized, pp. 329–345. Springer, Switzerland (2014)
9. Herman Parret, J.B.: Meaning and Understanding. Walter de Gruyer, New York (1981)
10. Kvanvig, J.: The Value of Knowledge and the Pursuit of Understanding. Cambridge University Press, Cambridge (2003)
11. Nivel, E., Thórisson, K.R., et al.: Bounded recursive self-improvement. RUTR 13006 (2012)
12. Nivel, E., Thórisson, K.R., et al.: Autonomous acquisition of natural language. In: IADIS International Conference on Intelligent Systems & Agents, pp. 58–66 (2014)

13. 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.: Bounded seed-AGI. In: Goertzel, B., Orseau, L., Snaider, J. (eds.) AGI 2014. LNCS, vol. 8598, pp. 85–96. Springer, Heidelberg (2014)
14. Pattee, H.H.: Evolving self-reference: matter, symbols, and semantic closure. In: Laws, Language and Life: Howard Pattee's Classic Papers on the Physics of Symbols with Contemporary Commentary, pp. 211–226 (2012)
15. Potter, V.G.: On Understanding Understanding: A Philosophy of Knowledge. Fordham University Press, New York (1994)
16. Kurzweil, R., Richards, J.W., Gilder, G.: Are We Spiritual Machines? Ray Kurzweil vs. the Critics of Strong AI. Discovery Institute Press, Seattle (2002)
17. Conant, R.C., Ross Ashby, W.: Every good regulator of a system must be a model of that system. Int. J Syst. Sci. **1**(2), 89–97 (1970)
18. Sloman, A.: What enables a machine to understand? In: Proceedings 9th International Joint Conference on AI, pp. 995–1001 (1985)
19. Steunebrink, B., Thórisson, K.R., Schmidhuber, J.: Growing recursive self-improvers. In: Proceedings of the 9th Conference on Artificial General Intelligence (2016)
20. Thórisson, K.R.: A new constructivist AI: from manual construction to self-constructive systems. In: Wang, P., Goertzel, B. (eds.) Theoretical Foundations of Artificial General Intelligence, pp. 145–171. Atlantis Press, Amsterdam (2012)