

Steve Austin Versus the Symbol Grounding Problem

John L. Taylor

Departments of Philosophy and Psychology
Humboldt State University
1 Harpst St. Arcata, California, United States of America
computationalphilosophy@hotmail.com

Scott A. Burgess

Department of Computing Science
Humboldt State University
1 Harpst St. Arcata, California, United States of America
sab43@humboldt.edu

Abstract

Harnad (1994) identifies the symbol grounding problem as central to his distinction between cognition and computation. To Harnad computation is merely the systematically interpretable manipulation of symbols, while cognition requires that these symbols have intrinsic meaning that is acquired through transducers that mediate between a cogitator and the environment. We present a careful analysis of the role of these transducers through the misadventures of Steve Austin, the Six Million Dollar Man. Putting Steve through a series of scenarios allows us to analyze what role transducers play in cognition.

Keywords: Symbol Grounding Problem, Transducers

1 Introduction

The symbol grounding problem is taken by some—including Harnad himself (Harnad 1994)—to bar the possibility of cognition being merely computation. In this paper we consider the role of transducers in cognition and whether the symbol grounding problem represents a legitimate problem for computationalism by asking and trying to answer the following questions: What exactly is a transducer? What role do transducers play in establishing meaning and enabling cognition? With the help of Steve Austin, the Six Million Dollar Man we explore several scenarios to answer these questions.

2 Key Definitions

Before outlining the symbol grounding problem, Harnad's use of several key concepts must be made clear. An analysis of computation and cognition requires a characterization of these processes and a means of determining when a candidate system is cogitating. Harnad attempts to fulfil these tasks with a definition of 'computer' and a refinement of the Turing Test.

Copyright © 2004, Australian Computer Society, Inc. This paper appeared at the *Computing and Philosophy Conference*, Canberra. Conferences in Research and Practice in Information Technology, Vol. 37. J. Weckert and Y. Al-Saggaf, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Computers are taken by Harnad to be symbolic systems, which is the standard, if not uncontroversial conception of computers. The conditions for something being a computer are threefold according to Harnad. Firstly, it must manipulate symbols. Secondly, it must be systematically interpretable. And, lastly, it must be implementation independent. Let us consider these conditions in greater depth (Harnad 1994).

Symbol manipulations (a.k.a. algorithms) are operations on symbols based solely on their shapes. Such manipulations are considered to be syntactic as opposed to semantic based on the meanings of symbols, rather than just their shapes. For example, Turing machines manipulate symbols on a tape according to a table of rules for transforming one symbol into another (See fig. 1). All that is required to do this is the capacity to distinguish symbols (presumably by their shapes) and treat them differentially.

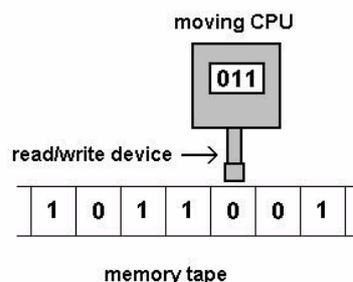


Fig. 1

Systematic interpretability (a.k.a. the cryptographers constraint) is the requirement that the symbol system should be decodable in a way that makes systematic sense. For instance, the actions of a particular Turing machine may be interpreted as subtracting two numbers. Harnad uses this condition to rule out the possibility of everything being a computer (contra Dietrich 2001).

Implementation independence is the condition that the shapes of the symbol tokens must be arbitrary in relation to what the symbols can be interpreted to mean. For instance, 'cat' does not resemble cats, nor does it require the presence of a cat to be tokened. Harnad considers the abstraction from physical details implied by

implementation independence to be significant in allowing computing machines to work at all. That is, if a computing device were dependent upon resemblances and

causal connections, its capacities would be severely limited.

Test	Symbolic Level	Sensory-motor Level	Neuro-molecular Level
T1: Partial Turing Test	Funct. equivalent symbolic I/O (one domain)	NA	NA
T2: Standard (Pen pal) Turing Test	Funct. equivalent symbolic I/O (all domains)	NA	NA
T3: Robotic Turing Test	Funct. equivalent symbolic I/O (all domains)	Funct. equivalent sensory-motor capacity	NA
T4: Total Turing Test	Funct. equivalent symbolic I/O (all domains)	Funct. equivalent sensory-motor capacity	Funct. equivalent neuro-molecular composition

Table 1: Turing Test Variation

Next, Harnad distinguishes four different levels (T1-T4) of the standard Turing Test to determine which level is appropriate for establishing whether or not a candidate system is a genuine cogitator. The Partial Turing Test (T1) requires that the candidate system be functionally indistinguishable from a person in its input and output regarding one domain of knowledge such as face recognition or chess. Many expert systems fulfill T1 for one domain or another. The Standard (Pen pal) Turing Test (T2) is familiar to all of us as the one Turing himself established (Turing 1964). The Robotic Turing Test (T3) requires that the system be functionally indistinguishable in sensorymotor capacity as well as being functionally indistinguishable in its symbolic input/output in every domain. The Total Turing Test (T4) requires everything that T3 does with the addition that it be functionally indistinguishable in its neuromolecular composition. Table 1 summarizes these distinctions.

3 The Symbol Grounding Problem

Given Harnad's definition of what constitutes a computer and his characterization of the "levels" of the Turing Test, an epistemic question naturally arises: What is the appropriate level for inferring the existence of a mind? It is clear that no one expects T1 to be sufficient for the task. T2, Harnad notes, is susceptible to criticism, including the Chinese Room Argument (Searle 1980, 1990, etc.) and the Simulation objection (Harnad 1994). The Chinese room, which is familiar to most of us, is supposed to illustrate the insufficiency of syntax to establish semantics, which implies that the standard Turing test (T2) fails to be the appropriate test for intelligence. The simulation argument as presented by Harnad reaches the same conclusion of T2's insufficiency:

A virtual plane does not really fly, a virtual furnace does not really heat, a virtual transducer does not really transduce; they are just symbol systems that are systematically interpretable as, respectively, flying, heating and transducing. All of this should be quite obvious. A bit less obvious is the equally valid fact that a virtual pen-pal does not think (or understand, or have a mind) -- because he is just a symbol system systematically

interpretable as if it were thinking (understanding, mentating) (Harnad 1994).

From these two arguments, Harnad concludes that T2 is not viable. Furthermore, Harnad (unlike Searle) finds T4 to be too strong a criterion for determining when a candidate system is a cogitator—which leaves T3. Harnad (1994) asserts that "T3 is the level at which we solve the "other minds" problem in everyday life, the one at which evolution operates..." To satisfy T3, Harnad argues, is to solve the symbol grounding problem (1990 a).

Consideration of this question leads Harnad to perceive a critical difference between computation and cognition.

I have no idea what my thoughts are, but there is one thing I can say for sure about them: They are thoughts about something, they are meaningful, and they are not about what they are about merely because they are systematically interpretable by you as being about what they are about. They are about them autonomously and directly, without any mediation (Harnad 1994).

The symbol grounding problem is the problem of connecting symbols to what they are about without the mediation of an external interpretation.

4 Harnad's Solution

How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their shapes, be grounded in anything but other meaningless symbols? To solve the symbol grounding problem, Harnad argues, symbolic capacities have to be grounded in robotic capacities. That is it, must be the case that:

...the robot itself can and does interact, autonomously and directly, with those very objects, events and states of affairs in a way that coheres with the interpretation. It tokens "cat" in the presence of a cat, just as we do, and "mat" in the presence of a mat, etc. And all this at a scale that is completely indistinguishable from the way we do it, not just with cats and mats, but with everything, present and absent, concrete and abstract. (Harnad 1994)

This has to be the case, Harnad reasons, because otherwise, the symbolic system would suffer from a kind of symbolic regress. The further use of symbols cannot ground the symbolic system any more than you can light a fire by piling the wood higher. Harnad offers an analogy to illustrate the problem. Imagine (assuming ignorance of Chinese) the futility of trying to learn Chinese with a Chinese-Chinese dictionary. Scanning meaningless symbols that reference other, equally meaningless, symbols leads nowhere. Similarly, he contends, just using symbols to give meanings to symbols is futile. By grounding our symbols in sensorimotor capacities, he avoids this regress. The process of establishing meaning terminates in the world itself, rather than endlessly cycling in the “hermeneutic circle” (Harnad 1990b). Guided by this view of meaning, Harnad has done considerable work with category perception and neural networks (e.g. Harnad 1995, 2001, etc.).

In solving the symbol grounding problem, we are forced outside of the definition of computation, according to Harnad: “But there is a price to be paid for grounding a symbol system: It is no longer just computational! At the very least, sensorimotor transduction is essential for robotic grounding, and transduction is not computation” (Harnad 1994). The difference between a computer and a cogitator is the difference between what can be described or interpreted as meaning *x* and what really is meaning *x*. A computer is just systematically interpretable as meaning *x*. Transduction is a necessary condition for really meaning *x* (and being a cogitator). Computers do not transduce. Thus, Harnad concludes, cognition is not the same as computation.

5 Enter Steve Austin: Defining the Role of Transducers

Harnad has elicited many objections to his argument: challenging his notion of implementation independence (Hayes 2001), his characterization of the Church-Turing Thesis and computationalism in general (Dietrich 2001), and not recognizing the full logical implications of accepting the Chinese Room argument (Searle 2001a, b), etc. However, he has not, to our knowledge, been challenged to refine his notion of transducer and clarify the role it plays in establishing meaning. Along these lines, several questions need to be answered. What exactly does Harnad mean by transducer? How does he (non-arbitrarily) distinguish transducers from a component of a computer (transistors, etc.) or the environment? How must transducers be attached to a symbolic system to establish meaning? What, exactly, is the relationship between transducers and intrinsic meaning?

With the help of Steve Austin (The Six Million Dollar Man) let’s explore scenarios of systematic replacement, deprivation and other situations to help us clarify what transducers are (as opposed to computer components and the environment), when semantics enters and exits the picture and whether any of these distinctions can be made in a non-arbitrary way. Table 2 enumerates these scenarios:

Variants of Steve	Transd.	Env.	History
Normal Steve/ Bionic Steve	Normal/ Artificial	Normal	Evolution/ Personal
Shut-in Steve/ Sensory Deprivation Steve	Normal	Indirect/ None	Evolution/ Personal
Disembodied Steve	None	Normal	Evolution/ Personal
Synthetic Steve/ Swamp Steve	Normal	Normal	None

Table 2: Scenarios

5.1 Scenario 1: Normal Steve/Bionic Steve

We begin our fanciful flight of conceptual analysis with a canonical cogitator—pre-accident Steve Austin. We presume that, as Harnad requires, he interacts with his environment via transducers (and, one presumes, effectors to query the environment with) and that his thoughts have meaning. What exactly is meant by transducer? A transducer is typically taken to be an electronic or organic device that converts energy from one form to another. Microphones, loudspeakers, thermometers, position and pressure sensors, and antennas are all electronic transducers; retinas, taste receptors (chemo-receptors) are organic transducers. This characterization of transducers is too vague to support the philosophical demands placed on it. Many objects in the environment and in the brain convert one form of energy to another (e.g. the photovoltaic effect and waste heat generated by neurons). Transducers may be distinguished from the environment and other components of a computer/brain by the fact that they do not merely transform and mediate energy—they also encode it. Transducers transform energy from the environment into a format that is manipulable by the symbolic system. This characterization of transducers may not satisfy everyone, but it is consistent with Harnad’s account and may allow us to consider further issues.

Additionally, a problem arises from the fact that typical PCs and of course those computers that attempt to pass the standard T2 Turing Test have transducers and effectors. The keyboard is a transducer, though an unglamorous one—the mechanical energy impinging on the keyboard (typically from fingers) is transduced into electrical signals. Both PC monitors and any sort of read-out on a T2 candidate computer are effectors. So, in what sense are Harnad’s T3 transducers different than T2? To answer this we may distinguish between symbolic and non-symbolic transducers. We may consider symbolic transducers transform symbolic inputs (like spoken or typed words) into encodings; non-symbolic transducers transform non-symbolic inputs (reflected light, chemicals) into encodings. Though we have doubts that arise from the potentially arbitrary or gerrymandered nature of this distinction, we will grant Harnad its

intelligibility so we may move beyond defining transducers to discerning their role in cogitation.

Lets rejoin Steve (not Harnad!) after his famous accident. Some of his transducers have been augmented or replaced altogether with artificial components. Thus far, we believe Harnad would not object to Steve being a cogitator—the substrate of Steve’s transducers does not matter, so long as they function as they did previously. Similarly, we may replace Steve’s brain (in part or *in toto*) with artificial components, so long as, like the transducers, the brain is functionally equivalent (Harnad 1994).

5.2 Scenario 2: Shut-in/Sensory Deprivation Steve

Suppose that post-accident Steve develops an acute case of agoraphobia. Consequently, he does not go outside and does not directly experience many things. Would his thoughts about things that he has not transduced lack meaning? The possibility that his thoughts are meaningless seems *prima facie* implausible—we have many meaningful thoughts about things we have not directly experienced all the time. If not, it would be a profound intellectual handicap; we would be unable to have meaningful thoughts about the things we learn from books, professors, friends, television and so on. So, unless we are willing to rule out meaning by indirect acquaintance (e.g. transmitted through pictures) and meaning by description (if you will), we (including Harnad) must suppose that grounding is commutative. That is, our thoughts may be meaningful by a grounding process that commutes through other people, books, television etc. This is reminiscent of Dretske’s Xerox principle (Dretske 1981).

To further illustrate the dissociation of transducers and meaning lets place our much-maligned Steve into a sensory deprivation chamber. Leaving only the possibility of symbolic transduction, can Steve cogitate? With the symbolic channel open, we may query Steve about his thoughts. It is counterintuitive to suppose that there is no mind present under these conditions. Presumably, if we were to take him out of the isolation chamber he would remember being in there, the queries and so on. It seems arbitrary to say that despite the seemingly intelligent way he may discuss what happened in the chamber that it lacks meaning, or that somehow meaning retroactively creeps into the thoughts he had in the isolation chamber. Even more difficulties are presented by simultaneously asserting the meaninglessness of his thoughts while in isolation and finding symbolically transduced thoughts meaningful. That is, we may have meaningful thoughts about things that we have not directly transduced, but have symbolically transduced, so why would Steve be incapable of having meaningful thoughts while in isolation? We must conclude that non-symbolic transducers need not be actively engaged for our thoughts to be meaningful. If this is the case, then perhaps merely being in possession of them is necessary for our thoughts to be meaningful?

5.3 Scenario 3: Disembodied Steve

Is the mere possession of transducers that would (counterfactually) function appropriately sufficient to ground Steve? To explore this question we strip Steve of his transducers and place his brain in a vat. Again we may communicate with his brain through symbolic means, but, in this case, no non-symbolic transducers remain. Can Steve cogitate?

It seems pretty obvious that he can cogitate, like in the case of the sensory deprivation chamber. It is arbitrary to maintain that his meanings would wink out of existence merely because his non-symbolic transducers are gone. If this were the case, then in the previous sensory deprivation scenario he would have intrinsic meanings until his (unused) transducers were removed, despite his functioning in the same way he was prior to the transducers being taken away and any protests he may make to the contrary. To illustrate the arbitrariness of stipulating that the possession of non-symbolic transducers is necessary for meaning consider that in the middle of a symbolic exchange we may take his non-symbolic transducers away and presumably the meanings would vanish from his thoughts mid-exchange, while having no effect on the exchange itself. In response to this objection, we may be tempted to say that it is significant that he once had transducers to ground him—perhaps meaning persists after some initial grounding through a personal developmental history or evolutionary history.

5.4 Scenario 4: Synthetic Steve

Is the fact that Steve once had transducers (and was thus grounded historically) sufficient to ground his symbols? If Steve’s history is taken to be significant to the meaningfulness of his thoughts, then we must embrace a form of semantic historicism, or even teleofunctionalism (Millikan 1984, etc.). Similarly, this stance is susceptible to Swampman-like thought experiments. Imagine that we make a molecular replica of Steve’s brain and immediately placed it into a vat with only symbolic transducers. Granting that Steve’s original brain had meaningful thoughts while in the vat and without non-symbolic transducers, why would the synthetic brain lack them? What is present in the original Steve Austin’s brain that is not in synthetic Steve? If synthetic Steve’s brain were implanted in the original Steve’s body he would function as the old Steve would. We may simply assert that intrinsic meaning occurs in the latter case but not the former, but it starts to look like intrinsic meaning is a difference that does not make a difference.

It may be maintained, despite these scenarios, that somewhere, at some time, there must have been transduction to establish meaning. In answer to this we may replace the synthetic Steve with a Swamp-Steve. That is, no thinking, transducing people were involved in the generation of a replica of Steve’s brain complete with symbolic apparatus for input and output. Again, why should this brain lack meaning?

6 Conclusion

Where is the difference that makes a difference in the role of transducers? It does not appear necessary to directly use non-symbolic transducers to have meaningful thoughts, nor is it necessary to be in possession of them at all; neither is it necessary to even have had them at one time. This leaves transducers with little to do in an account of meaning. It appears that either Harnad must adopt another theory of meaning having little to do with transducers (causal, teleosemantic or otherwise) or he can persist in emphasizing the role of transducers in fixing meaning despite the lack of conceptual necessity. If he chooses to adopt a new theory of meaning it may imperil his point about computation and cognition being different from one another and, of course, he inherits the various objections to these theories (the disjunction problem, misinformation, Swampmen, etc.). What then should we look to for an answer about intrinsic meaning?

Perhaps it is a better question to ask is what is the difference between intrinsic and extrinsic meaning and is this distinction both coherent and metaphysically significant? The notion of intrinsic meaning and its twin extrinsic meaning may be running Harnad's argument from behind the curtain. Intrinsic meaning is taken to be essential, whereas extrinsic meaning is imposed. For instance our thoughts about cats are taken by Harnad to be intrinsically meaningful, while the symbol string 'cat' is not. This may imply a form of privileged access, wherein our thoughts' meanings are known directly and without error. Without the intrinsic/extrinsic distinction and privileged access, both the need for T3 capacities and the Chinese Room argument dissolve. Intrinsic meaning and privileged access has been under attack by empirical studies as well as philosophical analysis for some time now and may be a (if not *the*) real point of divergence between computationalists and anticomputationalists.

Harnad is trying his best to accept Searle's intuition about syntax and semantics while being friendly to artificial intelligence. His third way (roboticism, if you will) is an ultimately unsuccessful attempt to navigate between anti-computationalism and computationalism. To both of these camps his approach is insufficient and for good reason—transducers do not seem to play a significant role in grounding meaning and the intuitions that run these arguments remain undisturbed. This leaves both computationalists and non-computationalists unmoved. Perhaps, then, we should back up and look at our primitive assumptions about intrinsic and extrinsic meanings and the notion of privileged access.

7 Bibliography

- Cangelosi, A., Greco, A. and Harnad, S. (2000): From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories. *Connection Science* **12**:143-162.
- Dennett, D. (1998): Can Machines Think. In *Brainchildren*. Cambridge, Mass.: MIT Press.
- Dietrich, E. (2001): The Ubiquity of Computation. *Psychology* **12**(040).
- Dretske, F. (1981): *Knowledge and the Flow of Information*. Cambridge Mass.: MIT Press.
- Harnad, S. (1990a): The Symbol Grounding Problem. *Physica D* **42**: 335-346.
- Harnad, S. (1990b) Against Computational Hermeneutics. (Invited commentary on Eric Dietrich's Computationalism) *Social Epistemology* **4**: 167-172.
- Harnad, S. (1992): The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *SIGART Bulletin* **3**(4) (October 1992) pp. 9 - 10.
- Harnad, S. (1994): Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. Special Issue on "What Is Computation" *Minds and Machines* **4**:379-390.
- Harnad, S. (1995): Does the Mind Piggy-Back on Robotic and Symbolic Capacity? In H. Morowitz (ed.). *The Mind, the Brain, and Complex Adaptive Systems*. Santa Fe Institute Studies in the Sciences of Complexity. Volume XXII. P. 204-220.
- Harnad, S. (2001): Grounding Symbols in the Analog World With Neural Nets -- a Hybrid Model. *Psychology* **12**(034) Symbolism Connectionism (1).
- Haugeland, J. (1985): *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT/Bradford.
- Hayes, P. (2001): Computers don't follow instructions. *Psychology* **12**(046).
- Millikan, R. G. (1984): *Language, Thought, and Other Biological Categories*. Cambridge, Massachusetts: MIT Press.
- Searle, J. (1980): Minds, brains and programs. *Behavioral and Brain Sciences* **3**: 417-424.
- Searle, J. (1990): Is the brain's mind a computer program?. *Scientific American* **262**: 26-31.
- Searle, J. (2001a): The Failures of Computationalism: i. *Psychology* **12**(060) Symbolism Connectionism (27).
- Searle, J. (2001b): The Failures of Computationalism: ii. *Psychology* **12**(062) Symbolism Connectionism (29).
- Turing, A. M. (1964): Computing machinery and intelligence. In *Minds and machines*. A. Anderson (ed.), Englewood Cliffs NJ: Prentice Hall.