Symbol grounding problem

Stevan Harnad (2007), Scholarpedia, 2(7):2373. doi:10.4249/scholarpedia.2373 revision #73220 [link to/cite this article]

- **Dr. Stevan Harnad**, Canada Research Chair, University of Quebec, Montreal, CANADA

The **Symbol Grounding Problem** is related to the problem of how words (symbols) get their meanings, and hence to the problem of what meaning itself really is. The problem of meaning is in turn related to the problem of consciousness, or how it is that mental states are meaningful. According to a widely held theory of cognition, "computationalism," cognition (i.e., thinking) is just a form of computation. But computation in turn is just formal symbol manipulation: symbols are manipulated according to rules that are based on the symbols' shapes, not their meanings. How are those symbols (e.g., the words in our heads) connected to the things they refer to? It cannot be through the mediation of an external interpreter's head, because that would lead to an infinite regress, just as my looking up the meanings of words in a (unilingual) dictionary of a language that I do not understand would lead to an infinite regress. The symbols in an autonomous hybrid symbolic+sensorimotor system -- a Turing-scale robot consisting of both a symbol system and a sensorimotor system that reliably connects its internal symbols to the external objects they refer to, so it can interact with them Turing-indistinguishably from the way a person does -- would be grounded. But whether its symbols would have meaning rather than just grounding is something that even the robotic Turing Test -- hence cognitive science itself -- cannot determine, or explain.

Contents

Words and Meanings

We know since Frege (http://plato.stanford.edu/entries/frege/) that the thing that a word refers to (i.e., its referent) is not the same as its meaning (or "sense"). This is most clearly illustrated using the proper names of concrete individuals, but it is also true of names of kinds of things and of abstract properties: (1) "Tony Blair," (2) "the UK's former prime minister," and (3) "Cheri Blair's husband" all have the same referent, but not the same meaning.

Some have suggested that the meaning of a (referring) word is the rule or features that one must use in order to successfully pick out its referent. In that respect, (2) and (3) come closer to wearing their meanings on their sleeves, because they are explicitly stating a rule for picking out their referents: "Find whoever is the UK's former PM, or whoever is Cheri's current husband". But that does not settle the matter, because there's still the problem of the meaning of the components of that rule ("UK," "former," "current," "PM," "Cheri," "husband"), and how to pick *them* out.

Perhaps "Tony Blair" (or better still, just "Tony") does not have this recursive component problem, because it points straight to its referent, but how? If the meaning is the rule for picking out the referent, what is that rule, when we come down to non-decomposable components like proper names of individuals (or names of *kinds*, as in "an unmarried man" is a "bachelor")?

It is probably unreasonable to expect us to know the rule (http://www.iep.utm.edu/w/wittgens.htm#H6) for picking out the intended referents of our words,-- to know it explicitly, at least. Our brains do need to have the "know-how" to *execute* the rule, whatever it happens to be: they need to be able to actually pick out the intended referents of our words, such as "Tony Blair" or "bachelor." But *we* do not need to know consciously *how* our brains do that; we needn't know the rule. We can leave it to cognitive science and neuroscience to find out how our brains do it, and then explain the rule to us explicitly.

The Means of Picking out Referents

So if we take a word's meaning to be the means of picking out its referent, then meanings are in our brains. That is meaning in the *narrow* sense. If we use "meaning" in a *wider* sense, then we may want to say that meanings include both the referents themselves and the means of picking them out. So if a word (say, "Tony-Blair") is located inside an entity (e.g., me) that can use the word and pick out its referent, then the word's wide meaning consists of both the means that that entity uses to pick out its referent, and the referent itself: a wide causal nexus between (1) a head, (2) a word inside it, (3) an object outside it, and (4) whatever "processing" is required in order to successfully connect the inner word to the outer object.

But what if the "entity" in which a word is located is not a head but a piece of paper (or screen)? What is its meaning then? Surely all the (referring) words on this page, for example, have meanings, just as they have referents.

Consciousness

Here is where the problem of consciousness rears its head. For there would be no connection at all between scratches on paper and any intended referents if there were no minds mediating those intentions, via their own internal means of picking out those intended referents.

So the meaning of a word on a page is "ungrounded." Nor would looking it up in a dictionary help: If I tried to look up the meaning of a word I did not understand in a (unilingual) dictionary of a language I did not already understand, I would just cycle endlessly from one meaningless definition to another. My search for meaning would be ungrounded. In contrast, the meaning of the words in my head -- the ones I *do* understand -- are "grounded" (by a means that cognitive neuroscience will eventually reveal to us). And that grounding of the meanings of the words in my head mediates between the words on any external page I read (and understand) and the external objects to which those words refer.

Computation

What about the meaning of a word inside a computer? Is it like the word on the page or like the word in my head? This is where the Symbol Grounding Problem (http://cogprints.org/0615/) comes in. Is a dynamic process transpiring in a computer more like the static paper page, or more like another dynamical system, the brain?

There is a school of thought according to which the computer is more like the brain -- or rather, the brain is more like the computer: According to this view (called "computationalism," a variety of functionalism (http://plato.stanford.edu/entries/functionalism/)), the future theory explaining how the brain picks out its referents (the theory that cognitive neuroscience will eventually arrive at) will be a purely computational one (Pylyshyn 1984). A computational theory is a theory at the software level. It is essentially a computer program: a set of rules for manipulating symbols. And software is "implementation-independent." That means that whatever it is that a program is doing, it will do the same thing no matter what hardware it is executed on. The physical details of the dynamical system implementing the computation are irrelevant to the computation itself, which is purely formal; any hardware that can run the computation will do, and all physical implementations of that particular computer program are equivalent, computationally.

The Turing Test

A computer can execute any computation. Hence once computationalism finds the right computer program, the same one that our brain is running when there is meaning transpiring in our heads, meaning will be transpiring in that computer too, when it is executing that program.

How will we know that we have the right computer program? It will have to be able to pass the Turing Test (TT) (Turing 1950 (http://cogprints.org/499/)). That means it will have to be capable of corresponding with any human being as a pen-pal, for a lifetime, without ever being in any way distinguishable from a real human pen-pal.

Searle's Chinese Room Argument

It was in order to show that computationalism is incorrect that Searle (1980) (http://www.bbsonline.org/documents/a/00/00/04/84/index.html) formulated his celebrated "Chinese Room Argument," in which he pointed out that if the Turing Test were conducted in Chinese, then he himself, Searle (who does not understand Chinese), could execute the very same program that the computer was executing without knowing what any of the words he was manipulating meant. So if there's no meaning going on inside Searle's head when he is implementing the program, then there's no meaning going on inside the computer when it is the one implementing the program either, computation being implementation-independent.

How does Searle know that there is no meaning going on in his head when he is executing the TT-passing program? Exactly the same way he knows whether there is or is not meaning going on inside his head under any other conditions: He *understands* the words of English, whereas the Chinese symbols that he is manipulating according to the program's rules mean nothing whatsoever to him (and there is no one else in in his head for them to mean anything to). The symbols that are coming in, being rulefully manipulated, and then being sent out by any implementation of the TT-passing computer program, whether Searle or a computer, are like the ungrounded words on a page, not the grounded words in a head.

Note that in pointing out that the Chinese words would be meaningless to him under those conditions, Searle has appealed to consciousness. Otherwise one could argue that there *would* be meaning going on in Searle's head under those conditions, but that Searle himself would simply not be conscious of it. That is called the "Systems Reply" (http://plato.stanford.edu/entries/chinese-room/#4.1) to Searle's Chinese Room Argument, and Searle rightly rejects (http://cogprints.org/4023/) the Systems Reply as being merely a reiteration, in the face of negative evidence, of the very thesis (computationalism) that is on trial in his thought-experiment: "Are words in a running computation like the ungrounded words on a page, meaningless without the mediation of brains, or are they like the grounded words in brains?"

In this either/or question, the (still undefined) word "ungrounded" has implicitly relied on the difference between inert words on a page and consciously meaningful words in our heads. And Searle is reminding us that under these conditions (the Chinese TT), the words in his head would not be consciously meaningful, hence they would still be as ungrounded as the inert words on a page.

So if Searle is right, that (1) both the words on a page and those in any running computer program (including a TT-passing computer program) are meaningless in and of themselves, and hence that (2) whatever it is that the brain is doing to generate meaning, it can't be just implementation-independent computation, then what *is* the brain doing to generate meaning (Harnad 2001a) (http://cogprints.org/4023/) ?

Formal Symbols

To answer this question we have to formulate the symbol grounding problem itself (Harnad 1990) (http://cogprints.org/3106/) :

First we have to define "symbol": A symbol is any object that is part of a *symbol system*. (The notion of single symbol in isolation is not a useful one.) Symbols are arbitrary in their shape. A symbol system is a set of symbols and syntactic rules for manipulating them on the basis of their shapes (not their meanings). The symbols are systematically interpretable as having meanings and referents, but their shape is arbitrary in relation to their meanings and the shape of their referents.

A numeral is as good an example as any: Numerals (e.g., "1," "2," "3,") are part of a symbol system (arithmetic) consisting of shape-based rules for combining the symbols into ruleful strings. "2" means what we mean by "two", but its shape in no way resembles, nor is it connected to, "two-ness." Yet the symbol system is systematically interpretable as making true statements about numbers (e.g. "1 + 1 = 2").

It is critical to understand the property that the symbol-manipulation rules are based on shape rather than meaning (the symbols are treated as primitive and undefined, insofar as the rules are concerned), yet the symbols and their ruleful combinations are all meaningfully interpretable. It should be evident in the case of formal arithmetic, that although the symbols make sense, that sense is in our heads and not in the symbol system. The numerals in a running desk calculator are as meaningless as the numerals on a page of hand-calculations. Only in our minds do they take on meaning (Harnad 1994).

This is not to deprecate the property of systematic interpretability: We select and design formal symbol systems (algorithms) precisely because we want to know and use their systematic properties; the systematic correspondence between scratches on paper and quantities in the universe is a remarkable and extremely powerful property. But it is not the same thing as meaning, which is a property of certain things going on in our heads.

Natural Language and the Language of Thought

Another symbol system is natural language (Fodor 1975). On paper, or in a computer, language too is just a formal symbol system, manipulable by rules based on the arbitrary shapes of words. But in the brain, meaningless strings of squiggles become meaningful thoughts. I am not going to be able to say what had to be added in the brain to make symbols meaningful, but I will suggest one property, and point to a second.

One property that the symbols on static paper or even in a dynamic computer lack that symbols in a brain possess is the capacity to pick out their referents. This is what we were discussing earlier, and it is what the hitherto undefined term "grounding" refers to. A symbol system alone, whether static or dynamic, cannot have this capacity (any more than a book can), because picking out referents is not just a computational (implementation-independent) property; it is a dynamical (implementation-dependent) property.

To be grounded, the symbol system would have to be augmented with nonsymbolic, sensorimotor capacities -- the capacity to interact autonomously with that world of objects, events, actions, properties and states that its symbols are systematically interpretable (by us) as referring to. It would have to be able to pick out the referents of its symbols, and its sensorimotor interactions with the world would have to fit coherently with the symbols' interpretations.

The symbols, in other words, need to be connected directly to (i.e., grounded in) their referents; the connection must not be dependent only on the connections made by the brains of external interpreters like us. Just the symbol system alone, without this capacity for direct grounding, is not a viable candidate for being whatever it is that is really going on in our brains when we think meaningful thoughts (Cangelosi & Harnad 2001).

Robotics and Categorization

The necessity of groundedness, in other words, takes us from the level of the pen-pal Turing Test, which is purely symbolic (computational), to the robotic Turing Test, which is hybrid symbolic/sensorimotor (Harnad 2000, 2007). Meaning is grounded in the robotic capacity to detect, categorize, identify, and act upon the things that words and sentences refer to (see entry for Categorical Perception).

To categorize is to do the right thing with the right *kind* of thing. The categorizer must be able to detect the sensorimotor features of the members of the category that reliably distinguish them from the nonmembers. These feature-detectors must either be inborn or learned. The learning can be based on trial and error induction, guided by feedback from the consequences of correct and incorrect categorization; or, in our own linguistic species, the learning can also be based on verbal descriptions or definitions. The description or definition of a new category, however, can only convey the category and ground its name if the words in the definition are themselves already grounded category names. So ultimately grounding has to be sensorimotor, to avoid infinite regress (Harnad 2005).

But if groundedness is a necessary condition for meaning, is it a sufficient one? Not necessarily, for it is possible that even a robot that could pass the Turing Test, "living" amongst the rest of us indistinguishably for a lifetime, would fail to have in its head what Searle has in his: It could be a Zombie, with no one home, feeling feelings, meaning meanings (Harnad 1995).

And that's the second property, consciousness, toward which I wish merely to point, rather than to suggest what its underlying mechanism and causal role might be. The problem of discovering the causal mechanism for successfully picking out the referent of a category name can in principle be solved by cognitive science. But the problem of explaining how consciousness can play an independent role in doing so is probably insoluble, except on pain of telekinetic dualism. Perhaps symbol grounding (i.e., robotic TT capacity) is enough to ensure that conscious meaning is present too, perhaps not. But in either case, there is no way we can hope to be any the wiser -- and that is Turing's methodological point (Harnad 2001b, 2003, 2006).

Note: *[this entry was published in Nature/Macmillan Encyclopedia of Cognitive Science; it has been revised and updated for Scholarpedia]*

References

Cangelosi, A. & Harnad, S. (2001) The Adaptive Advantage of Symbolic Theft Over Sensorimotor Toil: Grounding Language in Perceptual Categories. (http://cogprints.org/2036/) *Evolution of Communication* 4(1) 117-142.

Cangelosi, A.; Greco, A.; Harnad, S. From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. (http://cogprints.org/2132/) *Connection Science*12(2) 143-62.

Fodor, J. A. (1975) *The language of thought*. New York: Thomas Y. Crowell

Frege, G. (1952/1892). On sense and reference. In P. Geach and M. Black, Eds., *Translations of the Philosophical Writings of Gottlob Frege*. Oxford: Blackwell

Harnad, S. (1990) The Symbol Grounding Problem. (http://cogprints.org/3106/) *Physica D* 42: 335-346.

Harnad, S. (1994) Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. (http://cogprints.org/1592/) *Minds and Machines* 4:379-390 (Special Issue on "What Is Computation")

Harnad, S. (1995) Why and How We Are Not Zombies. (http://eprints.ecs.soton.ac.uk/3347/) *Journal of Consciousness Studies* 1: 164-167.

Harnad, S. (2000) Minds, Machines and Turing: The Indistinguishability of Indistinguishables (http://cogprints.org/2615/) . *Journal of Logic, Language, and Information* 9(4): 425-445. (Special Issue on "Alan Turing and Artificial Intelligence")

Harnad, S. (2001a) Minds, Machines and Searle II: What's Wrong and Right About Searle's Chinese Room Argument? (http://cogprints.org/4023/) In: M. Bishop & J. Preston (eds.) *Essays on Searle's Chinese Room Argument*. Oxford University Press.

Harnad, S. (2001b) No Easy Way Out. (http://cogprints.org/1624/) *The Sciences* 41(2) 36-42.

Harnad, S. (2003) Can a Machine Be Conscious? How?. (http://eprints.ecs.soton.ac.uk/7718/) *Journal of Consciousness Studies* 10(4-5): 69-75.

Harnad, S. (2005) To Cognize is to Categorize: Cognition is categorization. (http://eprints.ecs.soton.ac.uk/11725/) in Lefebvre, C. and Cohen, H., Eds. *Handbook of Categorization*. Elsevier.

Harnad, S. (2007) The Annotation Game: On Turing (1950) on Computing, Machinery and Intelligence. (http://eprints.ecs.soton.ac.uk/7741/) In: Epstein, Robert & Peters, Grace (Eds.) *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Kluwer

Harnad, S. (2006) Cohabitation: Computation at 70 Cognition at 20. (http://eprints.ecs.soton.ac.uk/12092/) In Dedrick, D., Eds. *Essays in Honour of Zenon Pylyshyn*.

Pylyshyn, Z. W. (1984) *Computation and cognition*. Cambridge MA: MIT/Bradford

Searle, John. R. (1980) Minds, brains, and programs. (http://www.bbsonline.org/documents/a/00/00/04/84/index.html) *Behavioral and Brain Sciences* 3(3): 417-457

Turing, A.M. (1950) Computing Machinery and Intelligence. (http://cogprints.ecs.soton.ac.uk/archive/00000499/) *Mind* 49 433-460 [Reprinted in *Minds and machines*. A. Anderson (ed.), Engelwood Cliffs NJ: Prentice Hall, 1964.]

Internal references

- Valentino Braitenberg (2007) Brain. Scholarpedia, 2(11):2918.
- James Meiss (2007) Dynamical systems. Scholarpedia, 2(2):1629.
- Walter J. Freeman (2007) Intentionality. Scholarpedia, 2(2):1337.
- Mark Aronoff (2007) Language. Scholarpedia, 2(5):3175.

Appendix 1

Brentano and the problem of "intentionality". Whenever there is a genuine problem but no solution, there is a tendency to paper it over with an excess of terminology: synonyms masquerading as important distinctions, variants tagged as if they were partial victories.

The "mind/body" problem is such a problem. It is a conceptual difficulty we have in equating and explaining "mental" states with "physical" states. (There is already the first hint of terminology multiplication here, with "mind/body" and "mental/physical".) "Mental" states also come under the guise of: "consciousness," "awareness," "subjectivity," "qualia," "intentionality," "1st-person states," and many other synonyms and paranyms.

"Intentionality" has been called the "mark of the mental" because of some observations by the philosopher Brentano (http://plato.stanford.edu/entries/brentano/) to the effect that mental states always have an inherent, intended (mental) object or content toward which they are "directed": I see something, want something, believe something, desire something, understand something, mean something etc.; and that something is always something I have *in mind*. Having a mental object is part of having anything in mind. Hence it is the mark of the mental. There are no "free-floating" mental states that do not also have a mental object. Even hallucinations and imaginings have an object, and even feeling depressed feels like something. Nor is the object the "external" physical object, when there is one. I may see a real chair, but the "intentional" object of my "intentional state" is the mental chair I have in mind. (Yet another term for intentionality has been "aboutness" or "representationality": thoughts are always *about* something; they are (mental) "representations" *of* something; but that something is what it is that the thinker has in mind, not whatever external object may or may not correspond to it.)

If this all sounds like skating over the surface of a problem rather than a real break-through, then the foregoing description has had its intended effect: No, the problem of intentionality is not the symbol grounding problem; nor is grounding symbols the solution to the problem of intentionality. The symbols inside an autonomous dynamical symbol system that is able to pass the robotic Turing Test are grounded, in that, unlike in the case of an ungrounded symbol system, they do not depend on the mediation of the mind of an external interpreter to connect them to the external objects that they are interpretable (by the interpreter) as being "about"; the connection is autonomous, direct, and unmediated. But *grounding is not meaning*. Grounding is an input/output performance function. Grounding connects the sensory inputs from external objects to internal symbols and states occurring within an autonomous sensorimotor system, guiding the system's resulting processing and output.

Meaning, in contrast, is something mental. But to try to put a halt to the name-game of proliferating nonexplanatory synonyms for the mind/body problem without solving it (or, worse, implying that there is more than one mind/body problem), let us cite just one more thing that requires no further explication: *feeling*. The only thing that distinguishes an internal state that merely has grounding from one that has meaning is that it *feels like something* to be in the meaning state, whereas it does not feel like anything to be in the merely grounded functional state. Grounding is a functional matter; feeling is a felt matter. And that is the real source of Brentano's vexed peekaboo relation between "intentionality" and its internal "intentional object": All mental states, in addition to being the functional states of an autonomous dynamical system, are also feeling states: Feelings are not merely "functed," as all other physical states are; feelings are also felt.

Hence feeling is the real mark of the mental. But the symbol grounding problem is not the same as the mind/body problem, let alone a solution to it. The mind/body problem is actually the feeling/function problem: Symbol-grounding touches only its functional component.

Supplementary References

Harnad, S. (1992) There Is Only One Mind/Body Problem (http://eprints.ecs.soton.ac.uk/6464/) . Symposium on the Perception of Intentionality, XXV World Congress of Psychology, Brussels, Belgium, July 1992 *International Journal of Psychology* 27: 521

Harnad, Stevan (2001a) Explaining the Mind: Problems, Problems (http://eprints.ecs.soton.ac.uk/5943/) . *The Sciences* 41: 36-42.

Harnad, Stevan (2001b) The Mind/Body Problem is the Feeling/Function Problem: Harnad on Dennett on Chalmers (http://cogprints.org/2130/) . Technical Report. Department of Electronics and Computer Sciences. University of Southampton.

See Also

Categorical Perception, Chinese Room Argument, Consciousness

Sponsored by: Anil Seth, University of Sussex, UK

Reviewed by (http://www.scholarpedia.org/w/index.php?title=Symbol_grounding_problem&oldid=12549) : Anonymous

Accepted on: 2007-05-06 20:09:07 GMT (http://www.scholarpedia.org/w/index.php?title=Symbol_grounding_problem&oldid=12549)

Category: Consciousness

This page was last modified on 30 January 2010, at 13:28.

This page has been accessed 93,118 times.