The Lot of the Casual Theory of Mental Content

Robert Cummins

The Journal of Philosophy, Vol. 94, No. 10 (Oct., 1997), 535-542.

Stable URL:
http://links.jstor.org/sici?sici=0022-362X%28199710%2994%3A10%3C535%3ATLOTCT%3E2.0.CO%3B2-N

THE LOT OF THE CAUSAL THEORY OF MENTAL CONTENT

The thesis of this paper is that the *causal theory of mental content* (hereafter CT) is incompatible with an elementary fact of perceptual psychology, namely, that the detection of distal properties generally requires the mediation of a "theory." I shall call this fact the *nontransducibility of distal properties* (hereafter NTDP). The argument proceeds in two stages. The burden of stage one is that, taken together, CT and the *language of thought hypothesis* (hereafter LOT) are incompatible with NTDP. The burden of stage two is that acceptance of CT requires acceptance of LOT as well. It follows that CT is incompatible with NTDP. I organize things in this way in part because it makes the argument easier to understand, and in part because the stage-two thesis—that CT entails LOT—has some independent interest and is therefore worth separating from the rest of the argument.

I. STAGE ONE: THE CONJUNCTION OF CT AND LOT IS INCOMPATIBLE WITH THE NONTRANSDUCIBILITY OF DISTAL PROPERTIES

Let us begin by clarifying some terms. By LOT, I mean the hypothesis that the human scheme of mental representation satisfies the following conditions:

(1) It has a finite number of semantically primitive expressions individuated syntactically.
(2) Every expression is a concatenation of the primitives.
(3) The content of any complex expression is a function of the contents of the primitives and the syntax of the whole expression.

The *locus classicus* of this thesis is Jerry Fodor's *The Language of Thought*.[1] I shall also use 'LOT' to refer to the language of thought itself rather than the hypothesis. The context will resolve any ambiguity this might otherwise introduce.

By CT, I mean the doctrine that the contents of the semantic primitives in the human scheme of mental representation are determined by their role in detection. The basic idea is just that the content of a primitive r in a system Σ is the property P if there is the right kind of causal connection between instantiations of P and tokenings of r by Σ's detectors. Causal theories come in a variety of fla-

[1] New York: Crowell, 1975.

0022-362X/97/9410/535–42

vors.[2] I am simply going to assume familiarity with these theories in what follows.[3] The only feature of the theory that will play a role in the discussion that follows is the following uncontroversial feature: assuming LOT, CT requires that every property represented by a primitive in LOT be detectable.

Now for NTDP. Distal properties generally cannot be directly transduced.[4] Instead, the detection of distal properties must be mediated by what we might as well call a theory about that property.[5] To detect cats (an instantiation of catness) requires a theory that says, in effect, what sorts of proximal stimuli are reliable indicators of catness. To detect cats visually, you have to know how cats look. The same goes for colors, shapes, and sizes: for these to be reliably detected visually under changes in perspective, lighting, and distance requires knowledge of such facts as that retinal-image size varies as the inverse square of the distance to the object.

Much of the knowledge[6] that mediates the detection of distal properties must be acquired: we are, perhaps, born with a tacit

[2] See Dennis Stampe, "Towards a Causal Theory of Linguistic Representation," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy* (Minneapolis: Minnesota UP, 1977), pp. 42-63; Fred Dretske, *Knowledge and the Flow of Information* (Cambridge: MIT, 1981); Fodor, *A Theory of Content and Other Essays* (Cambridge: MIT, 1990); Ruth Millikan, *Language, Thought, and Other Biological Categories* (Cambridge: MIT, 1984); David Papineau, *Reality and Representation* (Cambridge: Blackwell, 1987).

Some readers might wonder at the inclusion of adaptationist theories like those of Papineau and Millikan as flavors of CT. These theories differ from other causal theories only in using adaptational role to select those causal connections that are content fixing.

[3] A review can be found in my *Meaning and Mental Representation* (Cambridge: MIT, 1989).

[4] "Transduced" in the sense of Z. Pylyshyn, *Computation and Cognition* (Cambridge: MIT, 1984), chapter 6. The central idea is that transduction is a cognitively primitive operation that maps physical events at the sensory surface onto computational events "upstream." It is controversial whether there *are* any transducers in this sense in human functional architecture. It is not controversial that transducers thus conceived are not responsible for distal property detection.

This "engineering" conception of transduction is to be distinguished from that found in the mathematical theory of computation where a transducer is just any finite automaton.

[5] It is controversial in psychology just what the form and exact content the mediating theory might turn out to have. It might turn out to be a point in weight space—for example, Paul M. Churchland, *A Neurocomputational Perspective* (Cambridge: MIT, 1989), chapters 9-11. According to LOT, if it is learned, it will be a set of sentences in LOT. If it is innate, it might, in some sense, be implicit in the architecture. For a review of some ways such information might be implicit, see my "Inexplicit Information," in M. Brand and R.M. Harnish, eds., *The Representation of Knowledge and Belief* (Tucson: Arizona UP, 1986), pp. 116-26.

[6] I use 'knowledge' here as it is used in cognitive psychology, which does not require that what is known be true or justified. Knowledge in this sense is information that is treated as if it were reliable by the system that uses it. This is what allows us to say that the visual system must know such things as that retinal-image size varies as the inverse square of the distance to the object even though, in the ordinary sense of 'know', this was not widely known in Europe until the seventeenth century.

knowledge of Emmert's Law, but we are not born knowing how cats look, or with the ability to distinguish edges from shadows. We must, then, learn the theory that mediates cat recognition. Learning the theory will require formulating and confirming hypotheses such as these:

(A) Cats have whiskers.
(B) Cats have four legs.
(C) Cats have fur.

According to LOT, these hypotheses are represented as sentences in LOT. As such, they require for their formulation, a symbol for cats, that is, a |cat|.[7] But, according to CT, you cannot have a |cat| until you have the ability to detect cats. According to psychology, you cannot have the ability to detect cats until you have a theory of cats. According to LOT, you cannot have a theory of cats until you have |cat|s. So, you cannot make the conjunction of LOT and CT compatible with psychology.

Objections and replies. Objection one. That is all right for |cat|s, but what about |square|s? What reason is there to believe that we have to learn a theory of squares to recognize them?

Reply one. Suppose the objection is right about squares and circles and some other stuff. What are you going to do about cats? Your only option is to suppose that |cat| is not a primitive symbol in LOT. But if it is not a primitive symbol in LOT, then it must be a complex symbol in LOT, that is, a Tarskian combination of other symbols in LOT. This implies that |cat| can be defined in LOT in terms of symbols for properties whose detection does not require a learned theory. Good luck: this kind of reductionism has a dismal history; if you want to revive it, you are on your own.

Reply two. The only argument on offer for the idea that one *does not* have to learn a theory of squares to recognize squares is that you *could not* learn it. Indeed, Fodor argues quite correctly in *The Language of Thought* that you could not learn to recognize any property represented by a primitive of LOT, the argument being, in effect, just the one given above. But this cuts as heavily against LOT as it does against the idea that squares can be detected without the aid of acquired information.

Objection two. The thing is done by degrees. You start out with some relation between distal stimuli and *R*. Since you do not yet know much about cats, *R* cannot mean catness. It means, as it might be, *p*catness ('*p*' for proximal). Since your theories have to use the

[7] A |cat| is a mental representation whose content is the property of being a cat.

symbols you have got, you have hypotheses like these: *p*cats have whiskers;[8] *p*cats have four legs; *p*cats have fur. As the theory gets elaborated, eventually, we are able to recognize cats reliably.

Reply one. Why should the theory become elaborated? There is not, after all, an error signal that says, in effect, that the theory is not right yet. By hypothesis, the theory is, at any stage you like, as right as it could possibly be for *p*catness. This is because CT implies that you cannot have |*p*cat| in the theory unless you have a good enough theory of *p*cats to underwrite the connection between *p*cats and some symbol in LOT. To elaborate the theory in the right direction, you have to know that the theory is not yet a good enough theory of catness, and that, it would seem, requires having |cat| in your repertoire, applying them to various things—for example, *p*cats—and finding out that they are not cats.

It is easy to get confused about this, because one imagines someone who has not yet learned to recognize cats reliably looking at a cat and having the same percept as you and I would have. As George Berkeley would have said, he hath his eyes and the use of them as well as you and I; and surely that percept is a percept of a cat.

Quite right. But (a) it is not a |cat| since, as Berkeley also pointed out, it represents some particular cat of definite color, shape, and so on, whereas a |cat| represents any cat whatever; and (b) the structural similarities between the percept and the cat that make it so tempting (rightly, I would argue) to suppose it represents the cat before you are completely irrelevant to its representation content, according to CT. What matters, according to CT, are the causal relations between the representation and the property of being a cat; the intrinsic structural features of the representation are irrelevant. NTDP says nothing in your head has a chance of standing in the requisite causal relation to the property of being a cat unless you have a theory of cats. Lacking such a theory, nothing in your head can represent that property, perceptually or any other way.

Objection two reformulated. What you do is try to acquire a theory that will underwrite a mental representation that translates the public-language word for cats. While learning about cats, you are also learning about 'cat'. You say 'cat' whenever your detectors generate the symbol *R*. You keep refining the theory that mediates your detection system's tokening of *R* until you are saying 'cat' in the presence of the distal stimuli that prompt 'cat' in your language community.

[8] There is, of course, a corresponding problem for whisker detection and consequently for |whisker|s. So, perhaps, (1) should read, "*p*cats have *p*whiskers." We then work on |whisker| and |cat| simultaneously.

Reply one. This objection falsely implies that only humans (and other language users, if any) can mentally represent the property of being a cat. Cats recognize conspecifics, and hence, according to all but the most hopelessly chauvinistic version of LOT, cats represent catness.[9] Moreover, nonverbal children can recognize cats, as discrimination experiments show. Moreover, you can recognize things for which your language has no word. The objection under consideration will require that all of these be things for which you have a completely adequate paraphrase. This is surely an empirical question. Moreover, it is the sort of empirical question that a theory of content should leave open. Whatever the arguments are for LOT and CT, they should not settle issues like this.

Reply two. The objection currently on the table implies that the acquisition of 'cat' and |cat| are coeval. But is it an hypothesis dear to the hearts of LOTers that the acquisition of 'cat' is a matter of figuring out which mental representation best translates 'cat'. According to this view, learning the meaning of 'cat' is a matter of confirming the hypothesis that 'cat' means |cat| (that is, confirming '|"cat" means cat|'), and this is evidently ruled out by the scenario under consideration. Since I am not fond of the translation theory of language acquisition myself, I regard this as merely a polemical point, and I include it here only to sow the seeds of doubt in the opposition.

II. STAGE TWO: CT ENTAILS LOT

CT entails that mental representation is arbitrary in the following sense: any primitive representation could mean anything. You get this feature in CT for the following reason: if you can build a *P*-detector at all, you can arrange for it to output any symbol you like when it detects *P*.[10] So far as CT is concerned, then, any primitive symbol could mean any detectable property.

CT also entails that there are finitely many semantically primitive representations in the scheme of mental representation. To see this, begin with the observation that CT requires a detector for each primitive representation. Detectors do not have to be componential or functional units. All that is required is that, for each primitive rep-

[9] Maybe cats could get by with a |same species/variety as me|. Maybe. But (1) there is evidence that |me|s might be problematic for cats; and (2) in the context of the larger argument, acquiring a theory that will underwrite a |same species/variety as me| is going to be just as problematic as acquiring a theory that will underwrite |cat|s.

[10] It is tempting to suppose that the state of the detector when it detects *P* is the symbol in question, but this cannot be right, for then every time the system tokens that symbol, it will think it has detected *P*, and that will rule out thoughts like this: "Well, it is not a cat, but if it were, it would pounce on that mouse."

resentation, there is a mechanism/process that detects its target property. Since a finite system cannot incorporate infinitely many such mechanisms/processes, the question reduces to this: Can one mechanism/process detect infinitely many target properties?

It might seem so. Consider a planar array of photo-sensors. If we assume that a square pattern, regardless of location, size, or orientation on the array, is a representation of squareness, and make a comparable assumption about every shape, then this system will (depending on density, and discounting "viewing angles" other than those normal to the array) detect and represent infinitely many shapes. But notice that CT could not possibly underwrite such a representational scheme. For what we have done here is assume that any square pattern represents squareness regardless of its regular causes. To see this, notice that a particular pattern will only be caused by some squares, namely, those with the right location, orientation, and size-distance combination relative to the array. The causal theory will say that the pattern represents only those squares, not squareness. To fix this problem, we shall have to map each square pattern onto some single arbitrary symbol. To do that, we shall have to introduce an algorithm that abstracts from size, orientation, and location on the array. It will have to look for a closed pattern with four straight sides. To distinguish square patterns from trapezoids and parallelograms, it will have to test for right-angled corners; to weed out rectangles, it will have to test for side equality. It will, in short, have to know about squares. Ditto for any other shape that is represented by a primitive in the system. Since there is only a finite amount of memory available, this knowledge will have to be finite. So, the system can only map a finite number of shapes onto primitive representations.

It is clear that this argument generalizes. NTDP requires, in effect, a theory of *P* for the detection of *P*, and CT requires the detection of *P* as a precondition of the primitive representation of *P*. In a finite system, there is only so much room for theory, so it follows that there can, at any given time, be only a finite number of primitives in play. CT allows for a scheme in which the set of primitives is unbounded in the sense that, given enough memory, another can always be added.[11] But it does not allow for a scheme in which the number of primitives is infinite.

This may seem a pretty trivial result, but other stories about content do not have this consequence. As the shape example shows, if a representation represents everything structurally iso-

[11] Even this may be allowing too much, since, as every personal-computer owner knows, the architecture typically limits how much memory you can add.

morphic to it,[12] then it is possible to have a scheme with as many primitive shape representations as there are shapes (or something isomorphic to them) for the representations themselves to have. Finite precision will make it impossible for any actual system to exploit all that primitive representational power, but the scheme has it for all that.

Thus, CT implies that the scheme of mental representation has a finite number of arbitrary primitives. Moreover, since the scheme of mental representation needs to be productive, it follows that there must be a way of combining the semantically arbitrary primitives into complex representations such that the meaning of the complex is a function of the meaning of the constituents and their mode of combination. A scheme of finitely many semantically arbitrary primitives which allows for an unbounded set of semantically distinct complex representations whose contents are determined by their constituents and mode of combination is LOT as near as makes no difference.[13] So CT entails LOT.[14]

III. CONCLUSION

I have argued that, taken together, CT and LOT are incompatible with the nontransducibility of distal properties; and I have argued that CT entails LOT. It follows that CT is incompatible with the nontransducibility of distal properties. Since distal properties are not transducible, it follows that CT is false. Since the argument uses only a generic form of CT, it follows that CT is false in all its forms.

The conclusion can be generalized a bit. Nothing in the argument given here depends on the content-fixing relation's being causal; any covariationist semantics will face exactly the same problem, since (1) NTDP says that the only way to get reliable covariation between a mental state and a distal property is to have a theory of that property, and (2) covariational accounts make representation arbitrary in the sense required by the argument of stage two.

In *Representations, Targets, and Attitudes* (*op. cit.*), I argue that CT is a restricted form of functional-role semantics. CT is what you get

[12] See my *Representations, Targets, and Attitudes* (Cambridge: MIT, 1996).

[13] There is nothing in CT to guarantee that the combinatorics have to be Tarskian, but, then, neither must advocates of LOT insist on Tarskian combinatorics. It is just that Tarskian combinatorics are the only ones currently available for combining arbitrary symbols productively. LOT is Tarskian by default, not by inner necessity.

[14] Strictly speaking, what I have argued is that CT plus finite-detection resources plus productivity entails LOT. The finiteness of detection resources is, I suppose, an empirical premise, but it is uncontroversial. Productivity is an empirical premise as well, but it is a pretty secure premise, and it is a premise that both CTers and LOTers typically accept.

when you restrict the functional roles that determine content to the roles a representation plays in detection. The question therefore naturally arises as to whether functional-role semantics generally is incompatible with NTDP. I do not know the answer to that question. What is clear, however, is that nothing like the argument just rehearsed will work against functional-role semantics generally, since that argument turns essentially on facts about the detection of distal properties, and detection has no privileged role to play in functional-role theories of mental content.

ROBERT CUMMINS

University of Arizona