# Fodor on Causes of Mentalese Symbols

Tevfik Aytekin – Erdinç Sayan

*Bahçeşehir University, İstanbul*
*Middle East Technical University, Ankara*

**Abstract:** Jerry Fodor's causal theory of content is a well-known naturalistic attempt purporting to show that Brentano was wrong in supposing that physical states cannot possess meaning and reference. Fodor's theory contains two crucial elements: one is a notion of "asymmetric dependence between nomic relations," and the other is an assumption about the nature of the "causally operative properties" involved in the causation of mental tokens. Having dealt elsewhere with the problems Fodor's notion of asymmetric dependence poses, we show in this paper a difficulty with the other element of his theory concerning what kinds of properties are the causally operative ones in the tokenings of a semantic symbol in the brain of a perceiver. After presenting this difficulty, we examine three possible responses a Fodorian might make to our criticism.

**Keywords:** Causal theory of content, naturalistic semantics, asymmetric dependence, causal law, operative causes.

## 1 Introduction

Although it has gone through a number of revisions over the years, the core of Fodor's naturalistic theory of content, called the "asymmetric dependence theory" (ADT),[1] can be stated as follows:

[1] Fodor introduced his causal theory of content in Fodor (1987) and Fodor (1990). In his later writings he continued to defend a causal approach, such as in Fodor (1998, 12-15) and Fodor (2008, 196-220).

 

A symbol 'S' expresses the property *P* if:

(i) it is a law that instances of *P* cause tokenings of 'S,'[2]

(ii) sometimes tokenings of 'S' are lawfully caused by instances of non-*P*s,

(iii) non-*P*-caused 'S' tokenings asymmetrically depend on *P*-caused 'S' tokenings.[3]

ADT is designed by Fodor to explain, in purely naturalistic terms, what enables primitive Mentalese[4] symbols to acquire their semantic values (references). According to Fodor, Mentalese has a compositional semantics; that is, the meaning of a complex symbol is determined by the meanings of its constituents and how these constituents are put together. This reduces the problem of mental content, for Fodor, to that of explaining what determines the semantic values of Mentalese primitives. ADT enters the scene at this point . The essential feature of ADT is to derive the semantic properties of a primitive Mentalese symbol from the causal connections that the symbol has with the external world. This is done basically as follows: (i) and (ii) yield, based on the causal connections of the symbol with the world, a set of candidate properties to which the symbol might refer, and the asymmetric dependence condition (iii) selects one of them as the reference of the symbol.

## 2 The Problem

The part of ADT that has been a popular target of attack in the literature has been the notion of asymmetric dependence expressed in (iii).[5] It is not easy to spell out what it means for one causal connection to asymmetrically depend on another. We will mention one possible interpretation of that notion in the next section. However, in this paper

---

[2] Fodor himself uses the following condition in place of (i): There is a nomic connection between the property *P* and the property of *being a cause of 'S' tokenings*. But (i), as we stated above, seems simpler and is a version favored by many authors such as Baker (1991) and Antony – Levine (1991).

[3] As ADT makes it clear, Fodor takes causes to be properties.

[4] Mentalese is the hypothetical mental language, proposed by Fodor, in which thought processes take place.

[5] Loewer – Rey (1991) contains detailed discussions on the notion of asymmetric dependence. Also see Mendola (2003) and Rupert (2008) for some recent discussions of this notion.

we will not raise problems directly related to the notion of asymmetric dependence; instead we will direct our attention to clause (i), which we think deserves more attention than it gets in the literature.[6]

What the clause (i) asserts is that the relation between a symbol and its semantic content is grounded in a causal law having a certain form. According to ADT, a particular primitive Mentalese symbol[7] like HORSE, for example, refers to the property of *being a horse* because *horse* → HORSE[8] is a law and all other laws having the form *non-horse* → HORSE asymmetrically depend on the *horse* → HORSE law. As we have said, in this paper our focus will be on the kind of laws whose existence is required by the clause (i) of ADT, like the *horse* → HORSE law. The question we want to ask is this: Is it likely that such a law or nomic connection between *being a horse* and HORSE tokens exists as Fodor supposes?

Fodor seems to think that our reasons for believing in the existence of such a nomic connection are straightforward. He appears to think that the occurrence of thoughts about horses in the presence of horses is evidence enough for such a nomic connection: "… plant a horse right there in the foreground, turn the lights up, point the observer horsewards … and surely the thought 'horse there' will indeed occur to him." (Fodor 1987, 115) So, Fodor seems to believe that given the naturalistically specifiable background conditions such as good lighting, etc., the fact that horses cause horse thoughts in us shows the existence of a nomic connection between the property of *being a horse* and HORSE tokenings.

We agree that, given such background conditions, horses do cause horse thoughts (or HORSE Mentalese symbols). What we don't agree with is his assumption that in such cases the property of *being a horse* is the operative or causally responsible property. We want to argue that

---

6 In a previous paper, we have discussed some of the problems related to the asymmetric dependence condition (iii) and offered a theory of mental content that averts them. See Aytekin – Sayan (2010).

7 In what follows we will drop the prefix "primitive" and use the expression "Mentalese symbol" or simply "symbol" to mean "primitive Mentalese symbol."

8 We will use capital letters to denote Mentalese symbols and use italics to name properties. Here "*horse* → HORSE" is the short form of saying that "instances of the property of *being a horse* cause HORSE tokenings."

it is very unlikely that horses cause HORSE tokenings in virtue of the property of *being a horse* that horses instantiate, a property which Fodor takes to be a single, indivisible property.[9]

To make the point as clear as possible (if at the expense of sounding brutal) consider an ordinary horse which is placed sidewards in front of a perceiver. Given suitable conditions (such as sufficient lighting and appropriate distance), HORSE is tokened in the perceiver's head. Now, without letting the perceiver know about it, cut the horse vertically, carve its inside out, etc., and HORSE continues to be tokened in the perceiver's head. But it is clear that no horse as such remains there after that procedure, and hence no entity that possesses the property *being a horse*. This experiment strongly suggests that it is very unlikely that the property of *being a horse* is the operative property in those causal transactions. For, if carved horse halves can cause HORSE tokens too, it would clearly be a mistake to think that what is causally responsible or operative in the tokenings of HORSE is the property *being a horse*. But what Fodor needs to assure that HORSE tokens refer to horses is to pin down a law linking HORSE tokenings to *being a horse* as the operative property, and not to anything else.[10]

To emphasize the point we are making, consider another example: A photocell-controlled sliding door. When a human is close enough to the door, the door opens automatically. Does this fact show that there is a nomic connection between such door openings and the property of *being a human*? Intuitively, the answer is "No." It is plausible to think that the property that does the causing in this case is not the property of *being a human*, for non-human moving objects can also cause the door to open. It must be some other property, such as *being a moving object of a certain size*. Similarly, in the case of HORSE tokenings. When we token a HORSE in the presence of horses, it is highly implausible to think that the property of *being a horse* is what is causally responsible for that

---

[9] For a discussion of Fodor's choice of the unbroken property of *being a horse* (rather than a more elementary property like *having such and such a size and shape*) as the operative property in the causation of HORSE tokens, see Hattiangadi (2007, 134-140). Hattiangadi also discusses a number of problems with Fodor's choice of *being a horse* as the causally responsible property in tokenings of HORSE, which are different from the problem we present in this paper.

[10] Thus, for example, a causal law linking HORSE to *carved half horse* wouldn't do any good, even if there *be* such a law.

tokening. It must be some other property or properties instantiated by horses (and carved half horses).

If so, it would be wrong to infer to the existence of a *horse* → HORSE law from the fact that horses do cause HORSE tokenings.[11] But if Fodor's claim about the existence of such laws (laws like *horse* → HORSE or *cow* → COW) is unjustified, then his asymmetric dependence theory can't get off the ground, since that theory relies on clause (i) above, which is intended to assert the existence of such laws. In the case of horses, for example, Fodor intends (i) to state that it is a law that instances of the property of *being a horse* cause HORSE tokenings. In the following pages we will look at three possible responses a Fodorian might make to the criticism we just raised. But before that, let us briefly note that the other tool of the Fodorian apparatus, viz. asymmetric dependence, cannot be of any help to solve this problem. For asymmetric dependence to work there has to be already a nomic connection between the property of *being a horse* and HORSE tokenings. Asymmetric dependence is supposed to prevent certain causes of a symbol from entering into its content; its role is not to endow a symbol with content in the first place. According to Fodor's theory, only causation can provide content for a symbol. And our claim is that there is no nomic connection between the property of *being a horse* and HORSE tokenings to begin with.

## 3 Possible Responses

Let us now look at the first type of response. We have said that it is unjustified to infer the existence of a *horse* → HORSE law from the fact that horses cause HORSE tokenings, because the operative property cannot be taken to be the property of *being a horse* in those causal transactions. In the kind of examples we gave, only what we may call "surface properties," i.e. properties found in the exterior of horse bodies, are causally relevant. Causation of HORSE tokenings by superficial properties of horses (such as the property of *being (or looking like) a horse's side skin*) entails that non-horses too are capable of causing HORSE tokenings. This is a point Fodor concedes. Actually, according

[11] One might think that there might be some other way to justify the existence of the *horse* → HORSE law. Of course this is possible. But we are not aware of any other explicit justification given by Fodor.

to him, the notorious disjunction problem[12] is a result of the causation of 'S's by non-Ss, such as the causation of HORSE tokenings by cows or fake horses. Causation of HORSE tokenings by instances of the property of *looking like the side skin of a horse* does not show, so the Fodorian response might go, that there is no *horse* → HORSE law. It may be the case that both laws, namely, *horse* → HORSE and *side skin of a horse* → HORSE exist. If this is the case then there is no problem, Fodor would say, since the asymmetric dependence condition is designed to handle those kinds of situations. That is, since the *side skin of a horse* → HORSE law asymmetrically depends on the *horse* → HORSE law, HORSE expresses just the property of *being a horse*, according to ADT.

There is an important methodological question raised by this type of response: How do we know which properties are (or are not) nomically connected to which others? In particular, how do we know which properties are nomically connected to the symbol HORSE and which are not? While we think that this is an important question, a full consideration of it is beyond the scope of this paper. A combination of empirical tests and causal reasoning needs to be employed. We don't have a proof that the property of *being a horse* is not nomically connected with the symbol HORSE, but, as we have explained above, our point is that it is implausible to suppose so. On the other hand, Fodor writes as if it is obvious without giving any justification. (Except pointing to the ordinary cases of HORSE tokenings in the presence of horses, which fails, since, as we have argued above, it is highly unlikely that the property of *being a horse* is the operative property in those causal transactions.) And this makes Fodor's theory problematic and counterintuitive from the start.

The second response a Fodorian might want to give may be the following. Consider again the sliding door example. There is basically one

[12] 'The disjunction problem,' a term coined by Fodor, refers to a general difficulty for causal theories of mental content. Given that there may be a host of possible causes of a mental symbol, how can we find a principled distinction between those causes that are meaning constitutive (that is, expressed by the symbol) and those that are not? If we cannot find such a principled distinction, then a causal theory of content inevitably leads to the conclusion that a mental symbol expresses the disjunction of all of its possible causes. This is a problem because there are typically many causes of a mental symbol which are not meaning constitutive. For example, sometimes milk causes 'cow' tokens in us but surely 'cow' does not express milk.

property which the sensor attached to the door is sensitive to (viz. *being a moving object of a certain size*), but in the case of humans there are a myriad of ways a horse can cause a HORSE tokening in their heads. In other words, horses instantiate many properties which are capable of causing HORSE tokenings in us. Admittedly, the response continues, to say that there is a nomic connection between the property of *being a horse* and HORSE tokenings is a somewhat loose talk. Strictly speaking, what we have are nomic connections between certain properties instantiated by horses and HORSE tokenings. Such nomic connections provide a reliable link between the property of *being a horse* and the HORSE symbol. An this is what Fodor means by asserting the existence of a nomic connection between the property of *being a horse* and the HORSE symbol.

The line of reasoning involved in this second response seems very similar to Cram's (1992) interpretation of Fodor's theory of asymmetric dependence. In his explication of Fodor's notion of asymmetric dependence, Cram uses a diagram of the sort shown in the following figure:

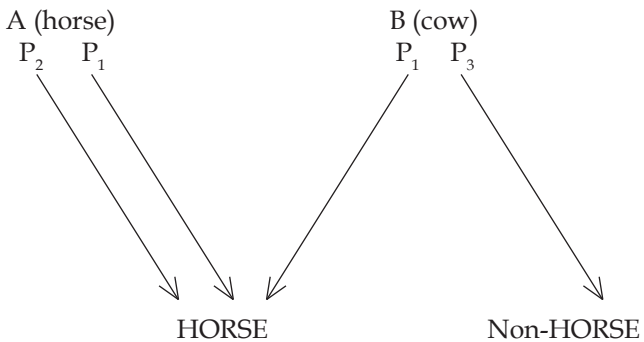


**Figure.** Cram's interpretation of asymmetric dependence

The figure shows two objects A and B (a horse and a cow) which can both cause a tokening of HORSE. We assume that each of these objects instantiates just two properties: A instantiates the properties $P_1$ and $P_2$, and B instantiates the properties $P_1$ and $P_3$. The arrows indicate the nomic or causal connections between the properties and the symbols. According to this picture, horses are capable of causing HORSE tokenings in virtue of instantiating the properties $P_1$ and $P_2$, each of which is

capable of causing HORSE tokenings in the absence of the other. Cows are also capable of causing HORSE tokenings in virtue of instantiating the property $P_1$.

It is now easy to see, according to Cram, how Fodor's notion of asymmetric dependence appears to solve the disjunction problem. Suppose that we "break" the causal connection between A and HORSE, which is to say that we are imagining that both of the laws $P_1 \rightarrow$ HORSE and $P_2 \rightarrow$ HORSE are broken, i.e. became inoperative. Since the individual B too can cause a HORSE tokening in virtue of its possessing $P_1$, the connection between B and HORSE would also get broken when the causal connection between A and HORSE is broken. But the converse is not true: even if we break the connection between B and HORSE, which is to say that the $P_1 \rightarrow$ HORSE law is disabled, since the causal route from property $P_2$ to HORSE remains intact, a connection between A and HORSE through $P_2$ is still present. This is why, explains Cram, the connection between B and HORSE is asymmetrically dependent on the connection between A and HORSE. And as a result, HORSE refers to *horse*, and not to *horse or cow*, even though cows too are capable of causing HORSE tokenings at times.

We think that Cram's interpretation makes Fodor's vexed notion of asymmetric dependence reasonably intelligible. Recall, however, that we are not interested, for the purposes of this paper, in the problems associated with the notion of asymmetric dependence. We have given Cram's interpretation as one possible way to cash out Fodor's allegation of a nomic connection between the property of *being a horse* and HORSE tokenings. According to this way of understanding Fodor's claim, to speak of a nomic connection between the property of *being a horse* and the HORSE symbol is just a shorthand for the nomic connections between certain properties horses instantiate and the HORSE symbol, as Cram's interpretation nicely illustrates.

But this kind of interpretation would create a major problem for Fodor's ADT. There is again a disjunction problem that occurs now at a different level. Referring back to our figure, one can claim that HORSE has a disjunctive content, namely, $P_1$ *or* $P_2$.[13] That is, the inter-

[13] This type of problem was also noted in Dretske (1994). In that paper Dretske argues that if an organism has multiple ways of detecting the presence of some substance—a situation similar to the one depicted in the figure above—then it can have a capacity for misrepresentation. However, in the same vein as the point we are making here, Dretske raises the worry

pretation of Fodor's theory à la Cram makes the notion of asymmetric dependence intelligible at the expense of reintroducing the disjunction problem, which the notion of asymmetric dependence was supposed to have solved. This problem seems to go unnoticed by Cram since he does not discuss it. Hence, a commendable interpretation as it may be of Fodor's alleged nomic connection between the property of *being a horse* and the HORSE symbol, it creates a major problem for Fodor's theory. For Fodor wants to say that HORSE refers to the property of *being a horse*, and not to disjunction of horse properties. As there seems to be no easy solution to this problem, we don't think that this option is a viable way to defend ADT against our criticism.

Finally, one might try to defend Fodor in the following third way. Fodor explicitly states that he is giving only sufficient, and not necessary and sufficient, conditions for intentional content:

> Don't forget, this stuff is supposed to be philosophy. In particular, it's an attempt to solve Brentano's problem by showing that there are naturalistically specifiable, and atomistic, sufficient conditions for a physical state to have an intentional content…. [S]olving Brentano's problem requires giving sufficient conditions for intentionality, not *necessary* and sufficient conditions. (Fodor, 1990, 96)

Accordingly, Fodor aims to solve this problem by providing the set of naturalistic conditions (i)-(iii) of ADT which he supposes to be sufficient for intentional content. But he does not have to show, according to this way of defending Fodor, that his conditions apply to humans; the "physical state" he is referring to may be a state of a robot or a computer, for example. So, even if ADT were not to apply to humans, this would not invalidate Fodor's theory. For example, imagine that future scientists have built a robot which can detect horses via the property of *being a horse* under certain background conditions. If it is possible to build such a robot, there turns out to be a nomic connection after all between the property of *being a horse* and a certain symbol which is tokened inside the robot, viz. the robot's HORSE token. If we accept the physical possibility of such a scenario, Fodor's claim about a nomic connection between the property of *being a horse* and some physical symbol is vindicated. If, in addition, this robot is capable of mistak-

---

that, under these circumstances, the internal state of the organism that gets caused in multiple ways can be said to indicate/mean a disjunctive condition rather than indicating/meaning the substance itself.

enly tokening HORSE in some other background conditions (so that the clause (ii) of ADT is satisfied), and furthermore, if the asymmetric dependence condition (iii) is also satisfied, then Fodor can assert that these robots have mental states which are about horses.

We would like to say two things about this type of response. First, we agree with Fodor that being able to give naturalistic sufficient conditions for intentional content is an important and difficult task. It is important because it relieves the physicalistic/materialistic worry about whether anything physical can have intentional content. So, even if Fodor's ADT does not apply to humans, if it does apply to robots, it certainly would have that merit. That is, given Fodor's general metaphysical aims, we do not think that his ADT theory should need to explain the intentionality of actual humans. For his main metaphysical purpose is to show that semantic/intentional properties are reducible to natural properties (those properties that figure in natural sciences). To serve this end, it would be enough for Fodor to show that some physical machine (not necessarily actual humans) can have intentionality. This would suffice to refute intentional irrealism (the view that there is no place for intentional properties in the natural world). Nevertheless, we think that Fodor's ADT would lose much of its interest if it failed to apply to humans. We should not forget that the naturalization project is part of a larger one for Fodor, namely, vindication of folk psychology. So, in fact the foremost challenge for him is to show within a naturalistic framework that *actual humans* have the intentional states we ascribe to them. A set of sufficient conditions valid for some physical entity would not do to vindicate folk psychology, unless this set of sufficient conditions also applied to actual humans. We don't think that intentional realists, including Fodor, would be happy if it turned out that there could be things with intentional states but actual humans were not among them.

Second, we don't think that Fodor would be more interested in applying his theory to hypothetical cases like the robot example than actual cases. For one thing, Fodor's ADT is itself a result of the consideration of these actual cases. Recall that the elementary causal theory Fodor initially considers and rejects as inadequate is dubbed by him "The Crude Causal Theory" (CCT):

> The Crude Causal Theory says, in effect, that a symbol expresses a property if it's nomologically necessary that *all* and *only* instances of the property cause tokenings of the symbol. There are problems

> with the "all" part (since not all horses actually do cause "horse" tokenings) and there are problems with the "only" part (cows sometimes cause "horse" tokenings; e.g., when they are mistaken for horses). (Fodor 1987, 100-101)

It is clear from the above words that the reason CCT is found to be problematic by Fodor is because it cannot deal with the actual cases, that is, those cases where some horses do not cause HORSE tokenings (such as when a horse is far away from the perceiver) and cases where other things (such as a cow) sometimes cause HORSE tokenings. (The latter case is in fact what creates the disjunction problem.) If we did not have to consider the actual cases, then we could even defend CCT as giving a sufficient condition for intentional content. Suppose that scientists have built a robot which detects only horses and is never misled by cows or other things. Isn't this physically possible? If it is, then there would be no disjunction problem for this robot. The CCT may have other problems as a naturalistic theory of intentional content[14] but if we do not consider the actual human cases, then there remains no disjunction problem that needs to be solved. So, we believe that, although giving sufficient conditions for intentional content is indeed a philosophically challenging task, we don't think this is the only thing that Fodor tries to achieve by formulating his ADT. Nor would a naturalistic theory of intentionality limited in its scope to some hypothetical nonhumans be much exciting for the rest of us.

## 4 Conclusion

Let us review the essential points of our argument against Fodor's theory. Fodor assumes that there are laws of the form *horse* → HORSE. We have argued that Fodor does not provide much of a justification for this assumption. He seems to think that it is obvious that there is a *horse* → HORSE law because horses cause HORSE tokenings.

---

[14] Such as the problem of "which link in the causal chain" that Fodor mentions in Fodor (2008, 205-207). Very briefly the problem is this. The causal connection between a mental symbol and its referent is typically a long chain of causes. There are a host of intermediate events in the causal chain between horses and HORSE symbols such as the production of retinal projections. Any causal theory of content needs to explain why the mental symbol HORSE expresses *being a horse* but not *being certain retinal projections*.

But to infer the existence of a *horse* → HORSE law from the mere fact that horses cause HORSE tokenings is unwarranted, since, as we argue, the property of *being a horse* is not likely to be the operative property in those causal transactions. And if there are no such laws, then ADT fails to apply to actual humans, to say the least. We have also argued against three possible attempts to defend Fodor's theory against our criticism. If what we claim is true, then Fodor's theory has a fundamental problem—owing to his clause (i)—which has been neglected for all these years.

*Tevfik Aytekin*
*Department of Computer Engineering*
*Bahçeşehir University*
*34353 İstanbul, Turkey*
*tevfik.aytekin@bahcesehir.edu.tr*

*Erdinç Sayan*
*Department of Philosophy*
*Middle East Technical University*
*06800 Ankara, Turkey*
*esayan@metu.edu.tr*

**References**

Antony, L. – Levine, J. (1991): *The Nomic and the Robust*. In: Loewer, B. – Rey, G. (eds.): *Meaning in Mind: Fodor and His Critics*. Oxford: Blackwell.

Aytekin, T. – Sayan, E. (2010): Misrepresentation and Robustness of Meaning. *Organon F* 17, No. 1, 21-38.

Baker, L. R. (1991): *Has Content Been Naturalized?* In: Loewer, B. – Rey, G. (eds.): *Meaning in Mind: Fodor and His Critics*. Oxford: Blackwell.

Cram, H. R. (1992): Fodor's Causal Theory of Representation. *Philosophical Quarterly* 42, 56-70.

Dretske, F. (1994): Misrepresentation. In: Stich, S. – Warfield, T. (eds.): *Mental Representation: A Reader*. Oxford, Basil Blackwell, 157-173.

Fodor, J. (1987): *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, MA: MIT Press.

Fodor, J. (1990): *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. (1998): *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Fodor, J. (2008): *LOT 2: The Language of Thought Revisited.* Oxford: Oxford University Press.

Hattiangadi, A. (2007): *Oughts and Thoughts: Rule-Following and the Normativity of Content*. New York: Oxford University Press.
Loewer, B. – Rey, G. (eds.) (1991): *Meaning in Mind: Fodor and His Critics*. Oxford: Blackwell.
Mendola, J. (2003): A Dilemma for Asymmetric Dependence. *Noûs* 37, No. 2, 232-257.
Rupert, R. (2008): Causal Theories of Mental Content. *Philosophy Compass* 3, 353-380.