An Argument Against Fodorian Inner Sentence Theories of Belief and Desire *

Adam Pautz
Comments welcome! adam.pautz@gmail.com

Abstract: One of Jerry Fodor's many seminal contributions to philosophy of mind was his inner sentence theory of belief and desire. To believe that p is to have a subpersonal inner sentence in one's "belief-box" that means that p, and to desire that q is to have a subpersonal inner sentence in one's "desire-box" that means that q. I will distinguish between two accounts of box-inclusion that exhaust the options: liberal and restrictive. I will show that both accounts have the mistaken implication that in certain cases there can be radical but "secret" changes in a subject's beliefs and desires. I will suggest that the correct moral to draw is that we should instead accept what Eric Schwitzgebel has called a "surface-level" theory of belief and desire.

One of Jerry Fodor's many seminal contributions to philosophy of mind was his *inner sentence* theory of belief and desire. To believe that *p* is to have a subpersonal inner sentence in one's "belief-box" that means that *p*, and to desire that *q* is to have a subpersonal inner sentence in one's "desire-box" that means that *q*. The metaphor of "boxes" is to be spelled out in functional terms, but the details are typically not supplied.[1]

I will develop an argument against the inner sentence theory. The argument takes the form of a dilemma. Briefly, I will distinguish between two possible accounts of box-inclusion: *liberal* and *restrictive*. I will show that liberal accounts deliver the mistaken verdict of *secret scrambling* in certain hypothetical cases. By contrast, restrictive accounts deliver the equally mistaken verdict of *secret losing* in those hypothetical cases. In fact, restrictive accounts are a non-starter because they deliver mistaken verdicts in actual cases.

I suggest that the moral of the argument is that we should instead accept what Schwitzgebel has called a "surface-level" theory of belief and desire. Examples include the *phenomenal-dispositional theory* (Schwitzgebel 2001, 2002, 2013) and a suitable *interpretationist theory* (e. g. Lewis 1974, Williams 2016, 2020). Like myself, Schwitzgebel develops an argument from cases against the inner sentence theory and for a surface-level approach to belief and desire (2001, 2002, 2013). My argument will differ from his and will, I

* I am grateful to . . . .

[1] See Fodor (1975, 1978, 1986, 1987), Field (1978) and Schiffer (1981). Fodor (1975) held that the inner sentence theory also applies to animals who do not speak a public language. See Beck (2017) for a recent discussion of this issue. I take the name "inner sentence theory" from Fodor 1978.

hope, contribute to the case against the inner sentence theory and for a surface-level approach.[2]

Those who hold on to the inner sentence theory in the face of my dilemma must identify which horn of the dilemma they favor. Even those willing to "bite the bullet" will find something of value in my argument. I hope that my argument may draw more attention to the important issue of the conditions for "box-inclusion".

In §1, I formulate an initial, liberal version of the inner sentence theory. In §2, I argue that it implies "secret scrambling" in certain cases. In §3, I argue that a restrictive of the inner sentence theory may avoid secret scrambling but it implies "secret losing" and is otherwise problematic. In §4, I formulate the official argument. In §5, I draw some morals.

1. A SIMPLE VERSION OF THE INNER SENTENCE THEORY

Fodor believed that each of us has a neurally-realized, subpersonal language of thought ("mentalese"). He argued for this hypothesis on the grounds that it provides the best explanation of the systematicity and the productivity of thought (see e. g. 1987, appendix). He held (1975, 66–79) that we do not ever experience or think about the hidden inner sentences of this language. And he proposed:

> [1] *The Basic Theory*. To believe that *p* is to have a subpersonal inner sentence in one's "belief-box" that means that *p*, and to desire that *q* is to have a subpersonal LOT sentence in one's "desire-box" that means that *q*.

This requires unpacking.

First, meaning. I take the inner sentence theory to be neutral on how exactly the meanings of inner sentences are initially determined. Fodor himself accepted an "asymmetric dependence" theory for subsentential mentalese expressions. Then the grammar (itself somehow physically-determined) fixes the meanings of whole inner sentences. Other views are possible. However, I will take the inner sentence to be committed to

> [2] *Inertia of meaning.* Once the meaning of an inner sentence is initially fixed by way of connections to the world, it tends to retain that meaning whenever it is tokened in the belief-box or the desire-box, even when it is temporarily severed from its normal connections to perceptual inputs and behavioral outputs (Fodor 1986, 12-15)

[2] Briefly, Schwitzgebel's main argument concerns "in-between believing". By contrast, mine involves "secret scrambling" and "secret losing". Moreover, Schwitzgebel's argument, my argument takes the form of a dilemma which engages with the issue of box-inclusion that his discussion neglects (see footnote 6 below for more on the differences between my argument and Schwitzgebel).

As an illustration, Fodor notes you might acquire the delusional, false belief that an inanimate item "is alive" and start to take care of it ("companion delusion"). Or again, Williamson (2007, 259) points out that revisionary metaphysician might withhold ordinary predicates ("is a mountain") from paradigm examples, and we take it that their "mountain"-beliefs are in error. Such errors require that the meaning of a mentalese term is not radically fickle; once it acquires its meaning it retains that meaning despite temporary aberrations in use.[3] True, holistic "inferential role theories" of meaning deny this; but Fodor staunchly opposed holism and considered content to be independent of inferential liaisons.

Next, what suffices for an inner sentence to be in your "belief-box" or "desire-box"? Fodor was mostly silent on this neglected question. In general, while much has been written about what it is for an internal sentence to mean that *p* ("psychosemantics"), there has been relatively little discussion of what it is for an inner sentence to be in the belief-box or the desire-box. One of my central aims is to bring it to the fore.

Before looking at this question, a clarification. My talk of a "belief-box" and a "desire-box" should not be taken too literally. The inner sentence theory is only committed to the claim that to believe that *p* is to stand in a functionally-characterizable relation *R* to an inner sentence that means that *p*. To say that a subject *x* has inner sentence *y* in her "belief-box" is just to say that *x* stands in relation *R* to inner sentence *y*. Understood this way, talk of a belief-box is not committed to a single, undifferentiated store (Quilty-Dunn and Mandelbaum 2018).

With this clarification out of the way, we can return to our question: what suffices for an inner sentence to be in your "belief-box" or "desire-box"? Fodor and Lepore (1992, 116) tentatively suggest one *sufficient condition* for box-inclusion: "beliefs, in the course of contributing to the etiology of behavior, interact with desires according to some decision theory or other". (See also Fodor 1987, 69 and Mandelbaum 2014, 82.) That is:

> [3] *One* of the sufficient conditions for "box-inclusion" is: *If* subpersonal inner sentences $b_1, b_2, \ldots$ and $d_1, d_2, \ldots$ typically (*ceteris paribus*) interact to cause the actions which, according to $b_1, b_2, \ldots$, will satisfy $d_1, d_2, \ldots$, *then* inner sentences $b_1, b_2, \ldots$ are "in the belief-box" and inner sentences $d_1, d_2, \ldots$ are "in the desire-box".

In fact, there is a simple argument for thinking that Fodorians should accept [3]. (i) Charlie the chimp uses a stick to get termites from a hole in a tree. He believes *if I put this stick down that hole I will get those termite-like things* and desires *I will get those termite-like things*. (ii) So Fodorians must say that the underlying inner sentences count as being in the chimp's belief-box and

[3] Fodor 1987, 93 allowed that *long-term* changes in use eventually result in changes in meaning.

desire-box. (iii) The Means-End condition [3], understood as a sufficient condition, delivers this result, and there is no alternative plausible sufficient condition that is satisfied in this primitive case.

Proponents of the inner sentence theory who reject Means-End condition [3] as a sufficient condition for box-inclusion must specify an *alternative plausible sufficient condition that is satisfied in this primitive case*. It's hard to see what that might be.

Notice that [3] is only the very weak claim that Means-End is *one* sufficient condition for box-inclusion. No doubt proponents of the inner sentence will add other sufficient conditions to [3], involving for instance being implicated in planning and reasoning, that are satisfied in the case of more sophisticated thinkers like ourselves (Field 1978, 13).

Let us, then, provisionally suppose the inner sentence theory to be committed to [1]-[3]. And let us call it a *liberal inner sentence theory*, since it doesn't place especially strong constraints (e. g. rationality constraints) on box-inclusion (I will say more about my sense of "liberal" in §3). It therefore fits with Fodor's own opposition to such constraints (1987, 88-89). I don't say Fodor himself definitely accepted it. But it's a good place to start.

A final point. There are different ways of understanding the inner sentence theory. As is well-known, Fodor (1987, chap. 1) takes it to be a *vindication* of commonsense realism about belief and desire. He is a staunch realist about beliefs and desires (contrary to "eliminative materialists"). He says (1987, 24) that a vindication of realism must agree with commonsense verdicts about our beliefs and desires in *core cases*. And he holds that the inner sentence theory is a vindication of realism in this sense (1987, 16, 24). In line with this, he is concerned to answer various alleged counterexamples to his account – alleged cases in which it yields counterintuitive verdicts about mental content (e. g. 1987, 92-93; 1990, 100-117). On a quite different way of understanding the inner sentence theory, it is neutral on common sense realism about belief and desire. In fact, it shouldn't be formulated as a theory about what it is to *believe* or *desire* something at all. Rather, it is merely the weak claim that there exist some interesting states – relations to inner sentences – that can be mentioned in the explanation of behavior.

I am interested in the adequacy of the inner sentence theory understood in the first way. I like to think that we believe things and desire things. Such facts are not fundamental. As Fodor memorably put it, "if intentionality is real, it must really be something else" (1987, 97). So I am interested the inner sentence theory as a proposal about how the facts about our beliefs and desires might be pinned down by more basic facts in core cases. (Compare how Kripke's (1980) causal theory of reference for proper names is a theory about such reference is pinned down by more basic facts.) My aim here is to evaluate the theory when understood in this way.

2. THE LIBERAL INNER SENTENCE THEORY IMPLIES SECRET SCRAMBLING

Suppose a dog bites your leg and you intentionally and vigorously attempt to pull your leg away. You desire that the pain go away and believe that doing this will achieve that. Many other "twisted" interpretations fit your behavior equally well but are incorrect: for instance, that you want your pain to *increase* and believe that moving your leg this way will achieve that (Lewis 1983, 374-375; 1986, 38; Williamson 2007, 253ff; Williams 2016 and 2020).

I will argue that inner sentence theory [1]-[3] of how content incorrectly predicts that, in certain cases, such a perverse (mis-)interpretation would actually momentarily become *true of you* (your beliefs and desires would actually *become scrambled*), while *there is absolutely no accessible, person-level change* in your experiences or dispositions to act under any circumstances.

Case 1: Color Case. Suppose that you are one of our *prelinguistic* hominid ancestors: you are smart but *you lack public (outer) language* (and so inner speech as well). (Focusing on an individual without a public language will help me keep the case simple.) Some scientists decide to perform a simple psychophysical experiment. You are withheld food. Then you are shown two color chips over and over; sometimes they have the same color and sometimes different colors. You are given cheese whenever you press on the button while the apparent colors are distinct and you get an electric shock whenever you press while the apparent colors are the same. You are then shown a red chip and a green chip a few times in a row, and press on the button. (See Figure 1.) Each time you believe *that the apparent colors are distinct* and *that if you press when they are distinct you will receive some cheese*, and you desire *that you receive some cheese*. This causally explains your behavior.

[Figure 1]

Figure 1: a red color chip and a green color chip

Fodorians will apply their inner sentence theory to this case (as well as the other cases to be presented). Fodor held that the inner sentence theory applies in "core cases". He said that those are cases in which there are "causally efficacious attitude tokenings"; his motto is "No Intentional Causation without Explicit Representation" (1987, p. 25). Fodorians will say that the Color Case is clearly a "core case" in this sense. It is an example of intentional causation in which your belief and desire cause your behavior.[4] So they will apply the

[4] There are some philosophers – for instance, "interpretationists" about beliefs and desires – who will generally deny beliefs and desires are inner states that are causally efficacious in the production of behaviour. Such philosophers might deny that in the Color case you belief and desire cause your behaviour. But I am interested in what *Fodorians* will say about the case. Fodorians believe in intentional causation. And they will say that the Color Case is a paradigmatic case of intentional causation.

inner sentence theory to this case. In fact, Fodor explicitly applies the inner sentence theory to such cases of prelinguistic belief and desire (1978, 512; 1987, 53).[5]

So, on the inner sentence theory, here is what is going on in these initial trials of the Color Case. Although you lack an outer, public language, there are in your belief-box two subpersonal *inner* sentences: "the apparent colors are distinct" and "if I press on the button when the apparent colors are distinct I will receive cheese"; and there is in your desire-box the subpersonal sentence "I will receive cheese". (Of course, in fact they are nothing like English.) They cause you to press on the button.

For the sake of argument, suppose that the inner sentence theory is correct in this case. In fact, let us assume that the *liberal* inner sentence theory [1]-[3] is correct. Recall that this consists of the *basic theory*, *semantic inertia* (inner sentences can be "moved around" while retaining their meanings) and the *means-end condition* as one sufficient condition for box-inclusion.

But now suppose that, in a final trial, you are shown the same red and green color chips. We stipulate that your first-person experiences, behavior and behavioral dispositions, and overall functional organization are exactly the same as in previous trials. The only difference is that, because of a malfunction between your "experience-box" and your "belief-box", they have a different subpersonal realization. Proponents of the liberal inner sentence theory would describe it as follows: what goes into your belief-box are the inner sentences "the apparent colors are the *same*" and "if I press the button while the apparent colors are the *same* I will receive cheese"; and as before "I will receive cheese" goes into your desire-box. (Recall that I am assuming semantic inertia, so that inner sentences can be "moved around" while retaining their meanings.) Your belief-box is momentarily "misspeaking" but, since the differences "cancel out", your behavioral dispositions (and your experiences) remain the same.

Two clarifications. (i) The example does *not* involve *all* of your LOT sentences being scrambled in some systematic way; it only requires *local* scrambling. (ii) Keep in mind that in this example you are meant to be one of our *pre-linguistic* hominids. *A fortiori*, in this example, when your inner sentences are scrambled, there will *not* be any differences in your *speech dispositions* (e. g. you will not suddenly be disposed to utter the English sentence "the apparent colors are *the same*"), nor any differences in your susceptibility to *semantic priming*, and so on. In general, I am just *stipulating* that, between the initial trials and the final trial, there are no differences in your personal-level experiences or outer dispositions under any circumstances (counterfactual conditional are also preserved). Since the case is hypothetical, I am free to make such a stipulation.

[5] See Camp 2009 and Beck 2017 for relevant discussion.

Given my stipulations about the case, the *liberal* inner sentence theory [1]-[3] entails *secret scrambling*. True, everything *from the inside* is the same as in previous trials. Look at Figure 1 again to help you imagine the situation from the first person. You have the same experience as of totally different colors (red and green), and this guides you in pressing the button. So, for all the world, it *seems* that you once again believe *that the apparent colors are distinct*. Nevertheless, the inner sentence theory [1]-[3] implies that you "really" now *secretly* and *irrationally* mistakenly believe that the apparent colors are *exactly the same*. For, by [3] Means-End, though the inner sentences ("the apparent colors *are the same*", etc.) are scrambled, they still count as being in your belief-box and desire-box, since they *interact with one another to cause your behavior in the manner distinctive of beliefs and desires* (compare Charlie the chimp with the stick). Further, by [2] Semantic Inertia, they retain their usual meanings, despite being temporarily aberrantly connected to the world (as in the examples from §1).[6] Finally, by [1] the Basic Theory, there has been a seismic and irrational change in your color belief. This is true despite the fact that *all your conscious experiences, and all your dispositions to engage in consciously-accessible behaviors, remain exactly the same*.

We will see in §3 that some Fodorians may wish to move to a *restrictive* form of the inner sentence theory that avoids the verdict of secret scrambling. All I am saying now is that, given my stipulations about the case, the *liberal* inner sentence theory [1]-[3] undeniably implies this verdict.[7]

This verdict is incorrect. Fundamentally, there is just a temporary *change in the neurally-realized vehicles*. Above I pretended they are English sentences,

[6] As noted in §1, inner sentence theorists need the inertia of meaning to handle actual cases of error due to "aberrant use". Inner sentences can be "moved around" while retaining their meanings. Consistency demands that they say that semantic inertia holds in the present case too. For instance, when "same apparent color" is aberrantly tokened in in your head in response to an experience as of *different* apparent colors (Figure 1), it still means *same apparent color*. Against this, proponents of a "holistic conceptual role theory" might say that that it suddenly now means *different apparent color*. On such a theory, any change in "use" might result in a change in content. But it's unclear how such a theory might handle error. And, in any case, it would be far from Fodor's own anti-holistic theory which is my target here; it would be closer to the kind of holistic "conceptual role theory" that Fodor staunchly opposed.

[7] To show that a version of the inner sentence theory implies strange verdicts about what you believe, it is not enough to describe a case where an aberrant internal sentence is tokened in your head. In addition, we must show that the inner sentence satisfies the conditions for belief-box-inclusion. I think that a drawback of Schwitzgebel's interesting discussion (2013, 83) is that he ignores this issue. Fodorians can respond by saying that in his cases the relevant aberrant sentences don't count as being in the belief box. Utilizing the Lewis-Williams-Williamson point about underdetermination, I have shown that in certain cases your inner sentences can indeed be scrambled *while at the same time satisfying the conditions for box-inclusion laid out by the liberal inner sentence theory [1]-[3].* And in §3 I will show that *restrictive* forms of the inner sentence theory face different but equally serious problems. The result is a novel dilemma for the Fodorian theory. This specific dilemma has not been presented in the literature.

but of course they are nothing like English. As it might be, the neural symbols @# are momentarily replaced by some other neural symbols &#. The interchange totally *preserves functional organization.* Whether your *belief* has changed is up for grabs. In my view, we should treat this as an example of *alternative realization of the same mental state*. Again, look at Figure 1, so that you can imagine being in this situation. In the fourth trial, as in the previous ones, the apparent colors (red and green) are obviously distinct to you. You attend to their difference. This causes you to press the button, and get the cheese. Given the totality of basic facts about this situation, the right verdict, once again, is that you believe the apparent colors to be *distinct,* despite the different neural realizer.[8] The liberal inner sentence theory [1]-[3] is mistaken.

Of course, it's always possible to "bite the bullet". Proponents of the liberal inner sentence theory [1]-[3] could just accept secret scrambling. For myself, I want to say that we *just know* this verdict to be incorrect. (Compare: we *just know* that you cannot make a beautiful painting become ugly just by changing its hidden microstructure, without making any changes to any of its surface-level, accessible features.) But I would be content with a weaker claim: this case provides at least *some* reason to reject the liberal inner sentence theory [1]-[3] and to seek an alternative theory that avoids secret scrambling.

Case 2: Taste Case. You are one of our prelinguistic hominid ancestors. You are starving, and you are given vanilla ice-cream for the first time (without knowing what it is). You believe that this white stuff tastes *sweet and good* and want this sweet, good-tasting stuff in your mouth. This causes you to devour it for five minutes until it is all gone – a case of intentional causation. Assume, for the sake of argument, a subpersonal language of thought. Mostly, everything is normal. However, starting at time *t*, there is a temporary, 10-second malfunction: while everything stays the same in your taste-system responsible for your total taste experience, further downstream you temporally undergo a neural re-organization that the inner sentence would describe as follows: the subpersonal sentence "this tastes *horribly bitter and disgusting*" is tokened in your belief-box and "I will have this *specific bitter, disgusting stuff in my mouth*" goes in your desire-box. During this short interval, your taste-experience of the ice-cream (at the level of affective as well as sensory phenomenology), as well as your disposition to stuff it in your mouth, remain exactly the same. However, the liberal inner sentence theory [1]-[3] implies that, for this short 10-second interval, you suddenly, and for no reason, secretly acquired a new, irrational and totally false belief about the ice-cream (it tastes horribly bitter and disgusting) as well as a crazy desire (to have this spe-

[8] One of my stipulations is that in this case – and the cases below – is that your *conscious experiences* remain the same. Fodor himself did not much discuss conscious experience (except to say that no one has the slightness idea about how material things can have it). Still, he of course recognized its reality. And this is all I need to run the cases.

cific disgusting stuff in your mouth). You momentarily cycle between a normal belief-desire psychology and a crazy one. But we must reject this verdict. By stipulation, your wonderful (sensory and affective) experience of the ice-cream stays exactly the same in phenomenal character, and you are disposed to stuff it in your mouth. *Throughout* you evidently believe it tastes *good* (not disgusting), and you want this *good* stuff in your mouth, even if these mental states are differently realized by symbols at the subpersonal neural level. At the very least, we have *some* reason to reject the verdict of secret scrambling in this case.[9]

Case 3: Large Number. You are a very primitive hunter-gatherer without public language. You can only visually recognize ("subitize") exact numbers up to about four or five. However, somehow you also have a *subpersonal* "number sense" akin to a form of clairvoyance. Maybe scientists implanted it in your brain; or maybe it is part of your idiosyncratic biological endowment. Oddly, your number sense is limited to detecting only one state of affairs: the presence of exactly 117 things behind your head. And you have in your mentalese lexicon only *one* representation of a large exact number, namely, "117". Occasionally, unknown to you, a collection of exactly 117 things is placed directly behind your head; and, thanks to your "number sense", when this happens, the subpersonal inner sentence "there are exactly 117 things behind my head" is stored in your belief-box. Many psychosemantic theories imply that it "represents" *that there are exactly 117 things behind my head*. However, you have *no experience* of this process: no experience of 117 things behind your head, and no experience of the inner sentence. Also suppose that, whenever there are exactly 117 things right behind your head, some sentences also go into your desire-box, like "when there are 117 things behind my head I will scratch my ear" and "when there are 117 things behind my head I will walk straight ahead". Together, these inner sentences cause you to scratch your ear and begin to walk forward.

The liberal inner sentence theory [1]-[3] implies that, on these occasions, you acquire a *belief* that is true iff *there are 117 things exactly behind you*, even though you are a primitive hunter-gatherer with no *person-level* means of representing large numbers (e. g. no public language representing large numbers) and no experiences of what his behind you. Fodor (1987, 89) conjectured that such "anti-holistic" verdicts might be derived from his theory, but he didn't provide the details. My example here shows such verdicts may indeed by derived from the liberal inner sentence theory [1]-[3]. Of course, proponents of

[9] Here is a sketch of a more extreme case. Williams (2016, 2020) describes an ingenious perverse "bubble" (mis)interpretation of all of a subject's beliefs and desires. The interpretation nevertheless preserves the subject's means-end rationality and "structural" rationality more generally. I think it could be shown that the liberal inner sentence theory [1]-[3] has the absurd implication that at twelve-noon today (say) the insane "bubble" interpretation could momentarily become *true of you* for a short period *without your noticing* or *changing at all in any of your experiences or behavioral dispositions* because of a hidden global permutation of your hidden LOT inscriptions.

the inner sentence theory might just accept the verdict that in this case you occasionally acquire a belief that is true iff there are 117 things exactly behind you. But I think we know this verdict to be incorrect in this case. After all, on these occasions, you have *no idea* of what is happening right behind your head. But then you don't *believe* that there are 117 things there, even if you have a *subpersonal representation* carrying this information. At the very least, we have more reason to reject this verdict than to accept the liberal inner sentence theory [1]-[3].

When a theory delivers many counterintuitive verdicts, at *some* point we should give it up. For example, a simple analysis of knowledge as justified true belief faces a host of counterexamples involve epistemic luck (Gettier 1963). And the description theory of reference faces the famous series of counterexamples described by Kripke (1980). If you agree that we should reject these theories because of their many counterintuitive consequences, shouldn't you also agree that we should reject the liberal inner sentence theory [1]-[3] because of its many counterintuitive consequences in the kinds of cases I have described? At the very least, I have given a novel reason to doubt the LIBERAL inner sentence theory [1]-[3].

3. COULD A PLAUSIBLE RESTRICTIVE INNER SENTENCE THEORY AVOID SECRET SCRAMBLING?

As I define the "inner sentence theory", it is at least committed to the basic theory [1] and semantic inertia [2]. The only moving part concerns *box-inclusion*. The inner sentence theory examined in the previous sections, which combines [1] and [2] with [3], is an example of:

> *Liberal Version.* The correct theory of box-inclusion is *liberal*, in the sense that it implies the aberrant inner sentences in the above examples count as being in your belief-box and desire-box.

This means that it implies secret and irrational belief-desire change in my examples.

At this point, other proponents of the inner sentence theory might give up the Liberal Version and move to:

> *Restrictive Version:* The correct theory of box-inclusion is *restrictive* (as it might be, incorporates "rationality constraints"), in the sense that it implies the scrambled inner sentences in the previous examples do *not* count as being in your belief-box and desire-box, even though they satisfy the means-end condition [3].

To illustrate, consider the Taste Case. Recall that there is a 10-second interval in your five-minute chow-down of the ice-cream when, by some glitch, the subpersonal inner sentences "the ice-cream tastes horribly bitter and disgusting" and "I will have this horribly bitter and disgusting stuff in my mouth" are tokened in your head. However, throughout you retain the same delicious, sweet taste-experience of the ice-cream and the disposition to stuff it in your mouth. According to the Restrictive Version, these aberrant inner sentences do *not* count as being in your belief-box and desire-box. The idea is that, although they satisfy the "means-end" condition (they interact with one another to cause your behavior in the manner distinctive of beliefs and desires), they fail to satisfy certain more restrictive constraints on box-inclusion.

For instance, on one natural version of the Restrictive Version, the aberrant inner sentence "this stuff tastes bitter and disgusting" fails be in your belief-box because it fails to satisfy the *Evidence Constraint*: if an inner sentence (e. g.) is radically incongruent with your experiential evidence (e. g. your taste-experience of the ice-cream), it cannot count as being in your *belief-box*. And the aberrant inner sentence "I will have this horribly bitter and disgusting stuff in my mouth" fails to satisfy the *Desire-the-Good Constraint:* if an inner sentence is incongruent with any sane set of intrinsic values, it cannot count as being in your *desire-box* (compare Lewis 1994, 427).

In this way, the Restrictive Version would avoid *secret scrambling:* the verdict that you suddenly *believe* that the ice-cream tastes horribly bitter and disgusting and that you suddenly have the perverse *desire* that you will have this horribly bitter and disgusting stuff in your mouth.[10]

[10] I have said that the textual evidence strongly suggests that Fodor probably would have accepted the Liberal Version, and would have accept its consequence of secret scrambling in my examples. What about contemporary followers of Fodor? Take, for instance, Quilty-Dunn and Mandelbaum's (2018) sophisticated defense of a Fodorian view. Concerning what it takes for an inner sentence to be in the "belief box", they write (2018: 2370), "beliefs are acquired *ballistically and automatically*, put subjects into a *negatively valenced* motivational state when encountering disconfirming evidence, will *increase in strength* over time . . ." (see also "The science of belief: a progress report"). Does this suggest a Liberal Version of the inner sentence theory which implies secret scrambling in some possible cases (the version suggested by Fodor's remarks), or does it suggest a Restrictive Version that avoids this implication?

There is some reason to think it implies the Restrictive Version. Consider, for instance, my Color Case. In the final trial, even though you experience distinct colors, the aberrant inner sentence "the apparent colors are the *same*" is tokened in your head and helps to control your behaviour, thereby satisfy the Means-End condition. It can be assumed to satisfy all of Quilty-Dunn and Mandelbaum's conditions, but for the condition about negative valence: for, while "the apparent colors are the *same*" is tokened in your head, you are presented with evidence that the colors are different, and yet you *don't* undergo a negatively valenced motivational state (for by stipulation your experience and behaviors are the same as in previous trials). So *if* Quilty-Dunn and Mandelbaum hold that the negative valence condition is a *strictly necessary condition* on belief, then their view implies that the aberrant sentence "the apparent colors are the *same*" does *not* count as being in your belief box. That is, they advo-

But the Restrictive Version implies *secret losing*. For it holds that, during the short interval, the aberrant inner sentence "this stuff tastes bitter and disgusting" doesn't count as being in your belief-box. And, we can suppose, there is no *other* inner sentence in your belief-box about the taste of the ice-cream. *A fortiori*, during this short interval, no sentence about the taste of the ice-cream is immediately entailed by any sentence in your belief-box. That is, no sentence about its taste is "implicitly stored" in your belief-box in the sense of Field 1978 and Lycan 1986. Therefore, the Restrictive Version implies that, during the short interval, even though there is no change in your wonderful and sweet taste-experience of the ice-cream or your tendency to stuff it in your mouth, you inexplicably and irrationally entirely lose *any* ("explicit" or "implicit") belief about the taste of this ice-cream. This is so even *just before and after you believe it tastes sweet and good* because just before and after "it tastes sweet and good" is in your belief-box.[11] Likewise, the Restrictive Version also implies that, during this short interval, you lack any *desire* about the ice-cream, *even though just before and after you desired to eat it.*

But these verdicts are *also* incorrect. *Throughout* you evidently believe it tastes *good*, without any brief interval of agnosticism. And *throughout* you desired to eat it.[12]

---

cate a Restrictive Version that avoids secret scrambling in this case. Maybe they could likewise avoid secret scrambling in the Taste Case. (I am not sure what they would say about the *Large Number Case*.)

But, on this interpretation of their view, it faces the problems I raised in §3 for any Restrictive Version. First, it implies *secret losing* in the Color Case and the Taste Case – which is just as bad as secret scrambling. Second, it is overly restrictive in other cases. Mentally ill people believe things but don't experience negative affect when presented with disconfirming evidence. Further, simple animals have beliefs but it is not plausible that they *all* experience negative affect when presented with disconfirming evidence. (Also, even if actual cases don't fill the bill, we can imagine wholly dispassionate believers.)

Perhaps they will say that the negative valence condition is *ceteris paribus* and doesn't apply in the case of the mentally-ill person because of malfunction. But in that case it also doesn't apply in the Color Case because there is malfunction in this case too. And then their view doesn't block the problematic verdict of secret scrambling after all.

[11] Here is an analogy. Suppose that first you believe it will rain tomorrow, then for some reason you momentarily lose that belief and become agnostic, and then finally you again believe it will rain tomorrow. On the inner sentence theory, this is grounded in the fact that first "it will rain tomorrow" is in your belief-box, then briefly it is not in your belief-box (and nor is any sentence which immediately entails anything about the weather tomorrow), and finally "it will rain tomorrow" is back in your belief-box. Given the Restrictive Version of the inner sentence theory, this is exactly what happens with the inner sentence "this stuff tastes sweet and wonderful" in my case. So it has the parallel implication in this case, namely that you momentarily lose any belief about the taste of the ice-cream while just before and after believing it to be sweet and good.

[12] A restrictive theory might also endorse a *Holistic Inference Constraint:* if an isolated LOT sentence (e. g. "there are 117 things behind me") tokened in your head doesn't tend to cause the tokening of *other* sentences that it inferentially supports (e. g. "there are more than 4 things behind me"), then it doesn't count as being in your *belief-box*. So, a restrictive theory

I have just argued that the Restrictive Version mistakenly entails *secret losing* in my examples. There is another argument against the Restrictive Version. The Restrictive Version says that the correct theory of box-inclusion is restrictive, in the sense that it implies that the inner sentences "this stuff tastes bitter and disgusting" and "I will have this horribly bitter and disgusting stuff in my mouth" do *not* count as being in your belief-box and desire-box, even though they satisfy the means-end condition for beliefs and desires. This requires that there are certain additional necessary conditions on box-inclusion that these aberrant sentences do *not* satisfy. The formulation of Restrictive Version is neutral on what they might be. But the most obvious candidates are the constraints mentioned above: the Evidence Constraint and the Desire the Good Constraint. The idea is that, since "this stuff tastes bitter and disgusting" is radically incongruent with your experiential evidence, and since "I will have this horribly bitter and disgusting stuff in my mouth" is radically incongruent with any sane any sane set of intrinsic values, they cannot count as being in your belief-box and desire-box. What *other* features of these aberrant sentences in this example might possibly preclude them from being in the belief-box and the desire-box? So, the Restrictive Version seems to require something like the Evidence Constraint and the Desire the Good Constraint.

However, there are numerous counterexamples to these constraints. As Fodor himself emphasized, even ordinary people without mental illness often have extremely "mad" beliefs and desires that go against their evidence and "mad" desires that go against any sane set of intrinsic values (1986; 1987, 88-89). Indeed, he says, "I accept - in fact, welcome - what amounts to the conclusion that people can believe things that are *arbitrarily* mad" (1987, 88). He explicitly rejects strong restrictions on box-inclusion.[13]

In short, the Restrictive Version holds that the aberrant sentences "this stuff tastes bitter and disgusting" and "I will have this horribly bitter and disgusting stuff in my mouth" do *not* count as being in your belief-box and desire-box, even though interact to cause your behavior in the manner distinctive of beliefs and desires. The trouble is that any constraints on box-inclusion that are strong enough to yield that verdict in this case will be *overly restrictive:* they will be undermined by the fact that people have weird beliefs and desires in *other* cases.

4. THE ARGUMENT FORMULATED

The argument can now be put in this way:

---

might avoid the verdict, in the Large Number Case, that you *believe* that there are exactly 117 things behind your head. However, Fodor himself, as a staunch opponent of holism, would never have accepted the Holistic Inference Constraint (see e. g. 1987, 89).

[13] See Dub (2015, 97, fn.5).

1. If the inner sentence theory is correct, then either the Liberal Version or the Restrictive Version is correct.
2. The Liberal Version implies *secret scrambling*, which is unacceptable (§2).
3. The Restrictive Version implies *secret losing* and in general is *overly restrictive*, which are also unacceptable results (§3).
4. So the inner sentence theory is mistaken.

Let me explain premise 1. I defined the inner sentence theory as any theory that endorses the *basic theory* and *semantic inertia*. The *Liberal Version* is simply defined as any version that combines the inner sentence theory with a liberal account of box-inclusion (such as the Means-Ends condition [3]), that is, one that *implies secret scrambling in my examples*. The *Restrictive Version* is defined as any version that combines the inner sentence theory with a restrictive account of box-inclusion, that is, one that *doesn't imply secret scrambling in these examples*. These exhaust the options, because any version of the inner sentence theory will either imply, or fail to imply, secret scrambling in my examples.

Therefore, proponents of the inner sentence theory have a choice. They must accept either the Liberal Version or the Restrictive Version.

How might proponents of the inner sentence respond to this argument? Should they accept the Liberal Version or the Restrictive Version?

One response would be for them is to accept the Restrictive Version but maintain (contrary to premise 3) that it is acceptable. I myself think that this response is a non-starter for the reasons given in the previous section.

I think Fodor himself would have accepted the Liberal Version, which implies secret scrambling, and would have insisted (contrary to premise 2) that secret scrambling is acceptable. As mentioned in the previous section, he said, "I accept - in fact, welcome - what amounts to the conclusion that people can believe things that are *arbitrarily* mad" (1987, 88). This suggests that he would have also said that their beliefs (and desires) can *change* in ways that are arbitrarily mad. And this in turn suggests a liberal view on which secret scrambling is possible. True, this goes against common sense. But Fodor also said that "a lot of what common sense believes about the attitudes must surely be false" (1987, 15).

Against the Liberal Version, I have already explained in §2 why I think we just *know* that in these cases your beliefs and desires do not secretly change in very irrational ways while all your surface-level experiences and behavioral dispositions remain the same. For instance, in the Taste Case, we just know that, while are having a wonderfully sweet taste experience of the ice-cream and greedily consuming it, it is not the case that you mostly believe it tastes sweet and wonderful but for a brief moment you secretly believe it tastes horribly bitter. Throughout you believe it tastes *sweet* and you want this *sweet* stuff in your mouth. Even if there is a momentary change in your subpersonal neural state, this is better regarded as a case of multiple realizability of the

same mental states. And, in the Large Number Case, we just *know* that you do not acquire a *belief* that is true iff *there are 117 things exactly behind you*, even though you are a primitive hunter-gatherer no experiences of what his behind you. You have no idea what is going on behind you! At the very least, these verdicts are *extremely counterintuitive*. And when a theory has many counterintuitive consequences, at *some* point it becomes reasonable to reject it.

We reject other philosophical theories when they deliver a number of counterintuitive verdicts about cases, such as the justified-true-belief analysis of knowledge and the description theory of reference. Consistency demands that we likewise reject the Liberal Version of the inner sentence theory.

In my view, then, *both* the Restrictive Version and the Liberal Version of the inner sentence theory are mistaken. Since they exhaust the options, we must reject the inner sentence theory altogether.

5. A POSSIBLE MORAL: A SURFACE-LEVEL THEORY OF BELIEF AND DESIRE

Even though I have focused on the inner sentence theory, the target of my argument may be more general. My dilemma arises for any theory that endorses two ideas: the *inner representation theory of belief and desire* (in terms of subpersonal representations in a belief-box and a desire-box) and *the inertia of meaning*. I have focused on the most popular and well-motivated version of this theory, the *inner sentence theory*, which claims that these inner representations have combinatorial structure. But the argument may apply equally to versions this theory which do not endorse the assumption of combinatorial structure.[14]

If my argument can be generalized in this way, what theory of belief could possibly avoid it? What theory of theory of belief and desire should we put in the place of the inner sentence theory of belief and desire?

In my view, it may be that where the inner sentence theory and kindred theories go wrong is in holding that what we believe and desire is fixed by hidden representations. Instead, perhaps we should hold that what we believe and desire supervenes on more "surface level" facts, such as experiences, dispositions to act, interactions with the environment, and acceptance of public language sentences. Two subjects (or the same subject at different times) who agree in the relevant "surface level" facts must agree in what they believe and desire. Examples of this kind of approach include the *phenomenal-dispositional theory* (Schwitzgebel 2002, 2013), a suitable *interpretationist theory* (e.

[14] For instance, in one version of David Lewis's "interpretationism", contents are in the first instance assigned to *repeatable brain states*. This form of the view is developed by Williams in a recent book (2020, 5). These brain states may then be "moved around" while retaining those contents ("semantic intertia"). Like Fodor's inner sentence theory, this version runs the risk of implying secret scrambling in atypical cases (compare Lewis 1980 on "mad pain"). And this is so whether or not it endorses the assumption of constituent structure.

g. Lewis 1974, Williams 2016, 2020), and theories appealing to *cognitive phenomenology* (Mendelovici 2018).[15] Because they only appeal to surface-level facts, such theories run no risk of allowing for "secret scrambling" or "secret losing" in which the surface-level facts remain the same but your beliefs and desires radically change.[16]

Of course, proponents of such surface-level views must say something about Fodor's justly famous motivations for positing a hidden, inner sentences with constituent structure, for instance, his arguments from systematicity and productivity (Fodor 1987, Quilty-Dunn and Mandelbaum 2018). How might a surface-level theory explain systematicity and productivity?

My aim in this essay has been to develop an argument against the inner sentence theory of belief and desire. Here is not the place to develop an alternative. But, briefly, I think that proponents of a surface-level theory can answer Fodor's challenge. In particular, I think we should replace Fodor's *inner* sentence theory of belief with an *outer sentence theory*. For individuals like ourselves with a public language, a central way of believing that *p* is by understanding and accepting an *outer, public-language* sentence that means that *p*. Since accepting an *outer* sentence that means that *p* (or being disposed to accept) is a surface-level fact, such an *outer sentence theory* of belief, unlike Fodor's *inner sentence theory*, runs no risk of allowing for secret scrambling. And such an outer sentence theory might also explain the systematicity and productivity of thought (Carruthers 1996).

Indeed, Fodor (1978) seriously considered such an outer sentence theory of belief (calling *Carnap's theory*) as a rival to his own inner sentence theory. One reason why Fodor ultimately rejected it in favor of his inner sentence theory is that individuals (e. g. our prelinguistic ancestors) without an outer language can have beliefs. But this particular problem is not decisive. Proponents of the outer sentence theory can answer it by accepting *belief-pluralism*

[15] Here is why I say a "suitable" version of Lewis's interpretationism. As noted in the previous footnote, the state-based version of Lewis's interpretationism defended by Williams (2010) may be subject to my argument about secret scrambling. However, as Williams notes (2020, 6), another possible version of Lewis's interpretationism (1974) holistically assigns beliefs and desires in the first instance to a *subject-at-a-time*, namely the ones that best fit "interpretative constraints" given her experiences and behavioral dispositions. Such a version may avoid secret scrambling.

[16] I have argued that Fodor's inner sentence theory is inconsistent with a surface level theory because in my cases it implies secret scrambling or secret losing even though all the "surface-level facts" remain the same. I also suggested that Fodor would accept that result. But suppose I am wrong. In fact, suppose that some Fodorian theory implies that, in *all* cases, if the surface-level facts are held fixed, then so are beliefs and desires. Then the Fodorian theory faces another problem: it becomes indistinguishable from a surface-level theory! **To distinguish their theory from surface-level theories, they must allow that beliefs and desires can vary while surface-level facts remain the same. And if they can vary while surface-level facts remain the same, why not in radical ways, as in my examples? Then we are back to the problem of secret scrambling and secret losing**.

(see e. g. Dennett 1978, 304-309; Bermudez 2003; and Speaks 2010, 234ff). For instance, maybe some kind of interpretationism is right for language-independent belief. And then, once individuals have an outer language, they have another way of believing that *p:* by understanding and accepting an outer sentence that means that *p*.[17]

References

Beck, J. 2017. Do non-human animals have a language of thought? In *Routledge Handbook of Animal Minds*. Ed. J. Beck and K. Andrews. London: Routledge Press.

Bermudez, J. 2003. *Thinking Without Words.* Oxford: Oxford University Press.

Braddon-Mitchell, D. & F. Jackson. 2006. *Philosophy of Mind and Cognition*. Blackwell, Oxford.

Camp, E. 2009. A language of baboon thought? In Robert Lurz (ed.), *Philosophy of Animal Minds*. New York: Cambridge University Press.

Curry, D. 2018. Beliefs as inner causes: the (lack of) evidence. *Philosophical Psychology* 31: 850-877.

Dennett, D. 1978. *Brainstorms.* Cambridge, MA: MIT Press.

Dub, R. 2015. The rationality assumption. In *Content and Consciousness Revisited: With Replies By Daniel Dennett*. *Studies in Mind and Brain Vol. 7*. Ed. Muñoz-Suárez, C. and De Brigard, F.

Field, H. 1978. Mental Representation. *Erkenntnis* 13: 9-61.

Fodor, J. 1986. Banish discontent. In *Language, Mind, and Logic*, ed. J. Butterfield, 1-23. Cambridge, UK: Cambridge University Press.

Fodor, J. 1987. *Psychosemantics.* Cambridge, MA: MIT Press.

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

Fodor, J. and LePore, E. 1992. *Holism: A Shopper's Guide*. Blackwell, Oxford.

Gettier, E., 1963. Is Justified True Belief Knowledge? *Analysis* 23: 121–123.

Kripke, S. 1980, *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Lewis, D. 1974. Radical interpretation. *Synthese* 27: 331-344.

Lewis, D. 1983. New work for a theory of universals. *Australasian Journal of Philosophy* 61: 343-376.

Lewis, D. 1980. Mad pain and Martian pain. In *Readings in the Philosophy of Psychology*, ed. Ned Block, 216-222. Harvard University Press: Cambridge, MA.

Lewis, D. 1994. Reduction of mind. In *A Companion to the Philosophy of Mind*, ed. Samuel Guttenplan, 412-431. London: Blackwell.

Lycan, W. 1986. Tacit belief. In R.J. Bogdan (ed.), *Belief: Form, content, and function*, Oxford: Clarendon, 61–82.

[17] Fodor (1978) had other objections to an outer sentence theory of belief. And proponents of anti-Fodorian, surface-level theories of belief and desire faces other issues. For example, there is the "Blockhead" problems. For a recent discussion, see Williams 2020, 26-28. There are also problems about the causal role of belief. In my view, all these problems are answerable, but again this is not the place to take them up. For an answer to Blockhead, see Jackson and Braddon-Mitchell 2006, 122. For an answer to the issue about causal role, see Lewis 1994, 428-429. See Curry 2018 for a quite different response.

Mandelbaum, E. 2014. Thinking is believing. *Inquiry: An Interdisciplinary Journal of Philosophy* 57: 55-96.

Mendelovici, Angela. 2018. *The Phenomenal Basis of Intentionality*. Oxford: Oxford University Press.

Mölder, B. 2010. *Mind Ascribed: An Elaboration and Defense of Interpretivism*. John Benjamins.

Quilty-Dunn, J. and E. Mandelbaum. 2018. Against dispositionalism: Belief in cognitive science. *Philosophical Studies* 175: 2353–2372.

Schiffer, S. 1981. Truth and the Theory of Content. In *Meaning and Understanding*, Herman Parret and Jacques Bouveresse, Berlin: Walter de Gruyter, 204–222.

Schwitzgebel, E. 2001. In-between believing. *Philosophical Quarterly* 51: 76–82.

Schwitzgebel, E. 2002. A phenomenal, dispositional account of belief. *Noûs* 36: 249-275.

Schwitzgebel, E. 2013. A dispositional approach to attitudes: thinking outside the belief box, in N. Nottelmann (ed.), *New essays on belief*, New York: Palgrave Macmillan, 75–99.

Speaks, J. 2010. Explaining the Disquotational Principle. *Canadian Journal of Philosophy* 40: 211-238.

Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Williams, J. Robert G. 2016. Representational Scepticism: The Bubble Puzzle. *Philosophical Perspectives* 30: 419–442.

Williams, J. Robert G. 2020. *The Metaphysics of Representation*. Oxford: Oxford University Press.

Williamson, T. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.