Holistic Theories of Content and Instability

Ryan Matthew Ferguson

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Arts
In
Philosophy

Tristram McPherson, Chair
Benjamin C. Jantzen
James C. Klagge

April 3, 2014
Blacksburg, VA

Keywords: representation, intentionality, content, holism

Holistic Theories of Content and Instability

Ryan Matthew Ferguson

ABSTRACT

In this paper, I will defend two methodological theses, one negative and one positive, about how to develop a holistic theory of content for mental representations that avoids a problem peculiar to *holistic* theories, viz., the *problem of content instability*. The relevant debate between holists and anti-holists has focused on whether this problem provides an in principle barrier to developing a plausible holistic theory. On this front, the holists have won; defenders of holistic theories have convincingly argued that the anti-holists do not have a cogent argument from the problem of content instability to the impossibility of developing a plausible holistic theory. However, beyond this, little has been said about how to develop a holistic theory that avoids the problematic consequences of content instability; all that has been established is that it appears to be, in principle, possible to do so. This paper should contribute to making progress in this area. The two theses I will defend are about how to generate *useful constraints* on holistic theories so that they avoid content instability. The *negative* thesis of this paper is that the strategy of generating constraints suggested by the holists' response to anti-holist arguments, viz., appealing to properties of theories' determination functions, is a non-starter. The *positive* thesis of this paper is that the best way to develop useful stability constraints is to appeal to the explanatory role(s) that representations play in cognitive science theories.

Acknowledgements

I would like to thank my committee members--Tristram McPherson, Benjamin Jantzen, and James Klagge--for all of their very helpful feedback and support in developing this paper.

Table of contents

Chapter 1: Holistic theories of content and the problem of content instability

In this paper, I will defend two methodological theses, one negative and one positive, about how to develop a holistic theory of content for mental representations that avoids a problem peculiar to *holistic* theories, viz., the *problem of content instability*. The relevant debate between holists and anti-holists has focused on whether this problem provides an in principle barrier to developing a plausible holistic theory. On this front, the holists have won; the anti-holists do *not* have a convincing argument from the problem of content instability to the impossibility of developing a plausible holistic theory. However, beyond this, little has been said about how to develop a holistic theory that avoids the problematic consequences of content instability; all that has been established is that it appears to be, in principle, possible to do so. This paper should contribute to making progress in this area; the two theses I will defend are about how to generate *useful constraints* on holistic theories so that they avoid content instability (i.e., stability constraints). Before I say more about these theses, however, I will briefly explain the problem of content instability and the holist/anti-holist debate surrounding it.

By a *theory of content*, I mean a theory that provides an answer to the question: 'In virtue of what is *r* a representation of *x* rather than a representation of *y*?' A theory of content is a theory of what (non-semantic) facts, properties, or relations *determine*[1] the semantic content of a representation. A theory of content for a system of representations *Q* specifies a) the content-determining properties of or relations between representations in *Q* and/or relations between representations and distal objects or properties; and b) the principles according to which the contents of representations in *Q* are determined by the properties or relations specified in (a) (cf. Pagin 1997 p. 13). All of the properties or relations specified in (a) for the representations in *Q* constitute the *total determination base* for *Q* and the properties or relations specified in (a) for a given representation *r* constitute the *determination base* for *r* (cf. Pagin 2006 p. 219). The "determining principles" in (b) can be represented as a function whose domain is the set of all

[1] I mean 'determine' in a *metaphysical* sense, not an epistemic sense. In other words, a representation *r* having content *c* depends upon (or *r* has *c* in virtue of) *r*'s having properties *P* or being in relations *R* with other representations in *Q* or with distal objects or properties. I will also use 'assignment', e.g., 'contents assigned to representations...' in this same metaphysical sense.

possible determination bases for a representation and whose range is the set of assignable contents for a representation. This is the theory's *determination function*.

It is difficult to articulate exactly what makes a theory of content *holistic*. Several, quite disparate theses have been called 'content holism' or 'meaning holism' (see Pagin 2006; Dresner 2012). Consequently, it is beyond the scope of this paper to attempt to provide necessary and sufficient conditions for holistic theories of content. Rather, I will point to a characteristic feature that holistic theories tend to share. According to holistic theories, contents of representations are determined by (*inter alia*) *all relations of a certain kind(s) between representations in a system*; there is no special, proper subset of these relations that are content-determining (cf. Dresner 2012).[2] For example, a holistic inferential role theory of content may hold that the content-determining relations of representations are *all* of the inferential relations between representations. In contrast, a localist inferential role theory may hold that only some special proper subset of inferential relations are content-determining.

The *problem of content instability* is that holistic theories of content (e.g., certain conceptual/functional role theories; see Block 1986; Harman 1987) are susceptible to objections of the following form; call them 'instability objections'. i) Holistic theory of content *T* entails a certain degree of content instability; ii) this degree of content instability entails that ordinary phenomena, e.g. disagreement, changes of mind, intentional generalizations, etc., are (at best) improbable; iii) but these ordinary phenomena are clearly not improbable, in fact they occur quite frequently; iv) therefore, *T* is false.

Some anti-holists have argued that this is a perfectly general problem for holistic theories; they argue that it is a consequence of the distinctive structure of such theories that they are susceptible to instability objections (see Fodor 1990 pp. ix-xi; Fodor & Lepore 1992 pp. 13-16). To see why, consider a toy holistic inferential role theory, which defines the determination base for a representation *r* as *all* of the inferential relations that *r* bears to other representations in *Q*. The set of all *r*'s inferential relations is quite unstable; it is improbable that any two time-slices of a person, let alone any two people, have mental representations that stand in *all* the

[2] We might also include in this characterization a property of the pattern of content-determination that holistic theories usually engender. Namely, for holistic theories, the contents of representations within a system are determined with a *high degree of interdependence*. That is, "the assignment of [content] to one or more [representations] constrains the assignment of [content] to others" (Pagin 2006 p. 221). The content of a representation *r* depends upon the contents of *s* and *t* and vice versa. See Pagin (1997; 2006) for more details.

same inferential relations (at least for most representations). Now, conjoin the instability of *r*'s determination base with **Change**, which is alleged to follow from all holistic theories.

Change Any change in the determination base for a representation *r* in *Q* changes the content of *r* and the contents of all *s* in *Q* such that *Rrs* or *Rsr*, where *R* is a content-determining relation.[3]

This conjunction (**Change** & *r*'s determination base is unstable) seems to entail that disagreement, changes of mind, and intentional generalizations[4] are (at best) improbable.[5] All of these ordinary phenomena require that different people (or time-slices of a person) have content-identical mental representations (although see fn. 5). But, **Change** entails that different people (or time-slices of a person) have content-identical mental representations only if they have mental representations that stand in *all* the same inferential relations. And since it is improbable

[3] In fact, **Change** is less extreme than what some have said follows from content holism. Something like **Change** is discussed in Block 1993 pp. 38-9; 1995 p. 152; Pagin 1997 p.23; Jackman 1999 p. 362. But, Fodor 1990 pp. ix-xi and Fodor & Lepore 1992 pp. 13-16 say something like the following follows from holism. *Any* change in the total determination base for *Q* changes the contents for *all* representations in *Q* (also see Lormand pp. 55-6; Pagin 2006 pp. 225, 227). Since this more extreme thesis implies **Change**, we can ignore it for the purposes of this paper.

[4] We might also add translation (Jackman 1999 pp. 361-2) and deductive inferences (Lormand 1996 p. 56) to the list of ordinary phenomena that are rendered improbable by **Change**.

[5] I say 'seems to entail' because it is somewhat controversial whether this conjunction does in fact imply these consequences (e.g., disagreement, change of mind, intentional generalizations are improbable). Some holists have responded to instability objections by denying that this implication holds. There are roughly two ways that holists have tried to support this claim. One way is to reject the idea that disagreement, intentional generalization, etc. require content *identity* across minds; rather, disagreement, etc. only require content *similarity*. So, although **Change** makes it improbable that people have representations with *identical* contents, relatively small differences between people's determination bases leaves them with representations with sufficiently *similar* contents for them to disagree, etc. (Block 1986 p. 629; Churchland 1986; Harman 1993). However, it is controversial whether holists can come up with a plausible notion of content similarity. Fodor & Lepore (1992 pp. 17-22, Ch. 7; 1993 pp. 679-682; 1999) and Churchland (1993, 1998) discuss this at length. Also see Tiffany (1999) and Pagin (2006 p. 227).

The other way to support this claim is to endorse a two-factor theory of content according to which there are two types or aspects of content: narrow content, which is determined holistically (by, say, conceptual role), and wide content, which is truth-conditional and externalist (i.e., it's meant to account for Twin-Earth/Burge cases). Under this two-factor theory, disagreement, intentional generalization, etc. would be understood in terms of the stable wide content, rather than the unstable, holistic narrow content (Block 1986; 1993 pp. 56-60). However, even if one had a plausible two-factor theory (e.g., it answers "the nasty question: *What keeps the two factors stuck together?*" (Fodor & Lepore 1992 p. 170)), it's unclear how, according to this theory, there can be robust intentional generalizations that quantify over the holistic, unstable narrow contents; the same instability issues just re-arise for narrow content. This is troublesome because part of the motivation for having narrow contents in the first place is to allow for certain kinds of psychological explanations, e.g., explaining why Oscar and his physical duplicate on Twin Earth behave similarly despite their beliefs, desires, etc. having different (wide) contents (Block 1986 p. 620; 1993 p. 59). It is beyond the scope of this paper to assess whether either of these responses to instability arguments actually succeeds.

that any two people have mental representations that stand in all the same inferential relations,[6] it is improbable that people disagree, change their minds, or are subsumed under intentional generalizations.

But, defenders of holistic theories have offered a convincing response to this general anti-holist argument. It has been noted by several authors (McLaughlin 1993 p. 656; Pagin 1997 pp. 23-5; Jackman 1999) that **Change** does *not* follow from all holistic theories. **Change** follows from holistic theories only if their determination functions are *one-one*. That is, only if, according to these theories, every assignable content corresponds to *exactly one* possible determination base for a representation.[7] This implies that if there is a change in a representation's determination base, then there is a change in the content of that representation. But, a holistic theory need not have a one-one determination function. A holistic theory can have a *many-one* determination function (call this 'the Many-One response'). The holist can claim that the content of a representation is determined by, say, the totality of its inferential relations, but that multiple, distinct sets of total inferential relations (i.e., multiple determination bases) assign the same content to a representation. Analogously, one's final grade for a class can be determined by one's grades on homework, quizzes, and exams, but multiple, distinct sets of grades on homework, quizzes, and exams can determine the same final grade; any change in, say, one's quiz grades does *not* entail a change in one's final grade (Jackman 1999 p. 363). **Change** does not follow from such holistic theories because according to these theories there *are* changes in determination bases that do *not* change the contents of representations. A difference (or change) in determination base does not imply a change in content; not all holistic theories entail **Change** and thus, not all holistic theories are susceptible to instability objections.

Thus, anti-holists have not provided a convincing, in principle objection to developing a plausible holistic theory of content. But, beyond this, little has been said about how to go about developing a holistic theory that avoids entailing content instability. What criteria do we use to evaluate theories in terms of the stability that they confer to content assignments? Are there any (additional) constraints on holistic theories that we can appeal to so that we can avoid instability objections? The strategy that is immediately suggested by the Many-One response is to appeal to properties of theories' determination functions. Specifically, if a theory has a many-one

[6] For more detailed explanations of why **Change** renders disagreement, etc. improbable see sources cited in fn. 5, esp. Fodor & Lepore pp. 13-16.
[7] A simple example of a one-one function is $f(x) = 2x$.

function, then it can avoid content instability. However, in Chapter 2, I will argue that this strategy is a non-starter; it is *not* the case that if a theory has a many-one function, then it can avoid content instability. This is the *negative* thesis of this paper. The *positive* thesis of this paper is that the best way to develop useful stability constraints is to appeal to the explanatory role(s) that representations play in cognitive science (cog sci) theories. In Chapter 3, I will give two reasons that support this thesis and in Chapter 4, I will outline and illustrate a rough procedure for extracting useful stability constraints from cog sci theories.

Chapter 2: Strongly many-one determination functions are insufficient for avoiding content instability

We're looking for guidance about how to develop a holistic theory of content that does not entail content instability. The Many-One response offers a seemingly promising stability criterion for holistic theories: If a holistic theory of content *T* has a many-one determination function, then *T* does not entail content instability. This stability criterion provides a straightforward, precise way of distinguishing the plausible holistic theories from the implausible ones vis-á-vis content stability. By looking at the *structural relationship* between determination bases and assignable contents encoded in properties of a theory's determination function (e.g., *being many-one*), we can ascertain whether that theory entails content instability.

However, as will be shown shortly, this strategy is a non-starter. A holistic theory's having a many-one determination function is *not* sufficient for avoiding content instability. In fact, a theory's having a many-one determination function is a rather unreliable indicator of content stability. This is true even on a strong reading of 'many-one function'. This suggests that the strategy for developing useful stability constraints that this criterion embodies--appealing to the structural relationship between determination bases and contents encoded in a theory's determination function--is wrongheaded.

According to the stability criterion suggested by the Many-One response, a holistic theory of content *H* does not entail content instability if *H*'s determination function *h* is many-one. Recall that a determination function's domain is the set of all possible determination bases for a representation (as defined by the relevant theory of content) and a determination function's range is the set of assignable contents for a representation. On a literal reading of 'many-one', *H*'s determination function *h* is many-one iff it is *not* the case that for all members of the domain[8] *w* and *v* if $h(w) = h(v)$, then $w = v$ (i.e., *h* is non-injective). That is, according to *H*, there is *at least one* assignable content that has more than one determination base mapping to it. But, this is a rather weak constraint on theories of content; it is clearly not sufficient for avoiding content instability.[9] The stronger, more plausible interpretation of 'many-one function' is a function for

[8] Of course, assume that the domain and range are non-empty.

[9] All that is required for *H*'s determination function *h* to be many-one under this reading is that there are two determination bases that map to the same content; even if all other determination base-to-content mappings are one-

which *all* assignable contents have multiple determination bases mapping to them. Call such a function 'strongly many-one'.[10] Now, here is the proposed stability criterion: If holistic theory of content *H*'s determination function *h* is strongly many-one, then *H* does not entail content instability.

The problem with this criterion and the strategy it embodies is that the relevant properties of a theory's determination function can be altered without thereby changing the content stability that that theory entails. This is because the members of the domain for a determination function (viz., determination bases) are *defined theory-internally*. Defining, and thus individuating, determination bases is part of what a theory of content does. Consequently, one can take virtually any theory with a one-one determination function and redefine the inputs to this function such that the function is many-one without thereby increasing the degree of content stability that that theory entails. The trick is, roughly, to take out the determination base individuating conditions *C* for the one-one theory, individuate the determination bases much more finely, and "smuggle" *C* into the determination function, thus making the function many-one. Since these two theories engender identical content-assignment changes under identical conditions (according to *C*), they entail the same degree of content stability. The difference is that one theory puts the relevant conditions in the determination bases, as it were, and has a one-one determination function whereas the other theory puts the relevant conditions in its many-one determination function (see Appendix A for a more detailed argument for this point).

So, since the inputs to the determination function (determination bases) are defined theory-internally, it is not very difficult to transform a theory with a one-one determination function into a theory with a strongly many-one determination function, and to do so without increasing the degree of stability it entails. Consequently, these properties of determination functions are not reliable indicators of content stability; the requisite covariance between many-one functions and avoiding content instability does not obtain.[11] This suggests that investigating

one, *h* is many-one. Say that *c* is the content that has two determination bases, d_1 and d_2, mapping to it. For all representations whose contents are not *c*, any change in their determination bases changes their contents and the contents of the representations they are *R*-related to, (where *R* is a content-determining relation) given that the contents of those representations aren't *c*. It's easy to see how an instability objection against such a theory would go.

[10] More precisely, *H*'s determination function *h* is strongly many-one iff for all members of the range *c* (contents) there are some members of the domain *w* and *v* (determination bases) such that $h(w) = c$ and $h(v) = c$ and $w \neq v$.

[11] It seems that, at best, we can say that if we have two theories T_1 and T_2, which are identical in all respects (e.g., definitions of determination bases) except that T_1 has a one-one determination function and T_2 has a strongly many-

theories' structural relationships between determination bases and assignable contents is *not* a good strategy for developing useful stability constraints. We should doubt that properties of these relationships reliably track content stability.

I propose that we adopt an alternative strategy for developing useful stability constraints for holistic theories. Rather than looking at structural properties of theories, we should look for more precise, theory-independent[12] standards of content stability and develop constraints with respect to those standards. Specifically, we should appeal to the explanatory role(s) of representations in cognitive science theories. In the next chapter, I will provide two reasons for thinking that this is the best way to develop useful stability constraints.

one determination function, then T_2 will entail a greater degree of stability. That is, *ceteris paribus* theories with strongly many-one determination functions entail a greater degree of stability.

[12] That is, independent of *theories of content*.

Chapter 3: Two reasons to appeal to the explanatory role(s) of representations in cognitive science theories

The thesis of this chapter is that the best way to develop useful stability constraints is to appeal to the explanatory role(s) of representations in cog sci theories. Two lines of thought support this thesis.

One: Cog sci theories provide the most successful explanations of behavior in which contentful representations figure as explanans.[13] This motivates the following constraint on theories of content: A theory of content is adequate only if it is compatible with the explanatory role of representations in successful cog sci theories. That is, if representations, whose contents are determined according to theory of content *T*, are unable to fulfill their requisite explanatory roles in successful cog sci theories, then we have reason to reject *T*. This constraint has consequences for the degrees of content stability that theories of content are required to entail; in order for representations to play their explanatory roles, their contents must remain sufficiently stable when exposed to perturbations specified by the relevant cog sci theories (e.g., formal procedures performed on representations).

The details of this will be provided in the next chapter. The salient point for this chapter is that we have a clear and compelling *motivation* for looking at the explanatory role(s) of representations in service of developing stability constraints. Namely, in order for representations to "be their best" as explanatory posits, they need to fulfill the requisite explanatory roles in successful cog sci theories. And, conveniently, in order for representations to fulfill these explanatory roles, they must meet certain content stability requirements.

Two: Perhaps the only alternative, prima facie plausible resource for developing stability constraints is our pre-theoretic, folk content-attributions. We could exploit this resource to develop stability constraints like this: Determine when and how often disagreement, changes of mind, etc. occur according to a given theory of content. Then, appeal to our folk content-

[13] Of course, it is controversial whether and to what extent posited contentful representations contribute to the explanatory success of (certain kinds of) cog sci theories. See e.g., Stich (1983); Chomsky (1995); Ramsey (1997; 2007). See Egan (2012) and Pitt (2012) for discussion. And, of course, not all cog sci theories posit representations as explanans, e.g., dynamical systems theories.

attributions as evidence for when and how often disagreement, changes of mind, etc. *actually* occur and formulate stability constraints according to these findings.[14]

However, if we want *useful* stability constraints--constraints that do more than rule out obviously implausible theories--we can see that our folk content-attributions are ill suited for this purpose. The reason is that the set of cases in which our folk content-attributions provide clear answers is severely limited. Our intuitions are clear in extreme, obvious cases, but when we consider more moderate cases, the cases that are important for obtaining useful information about stability constraints, our intuitions are vague or absent. Even if we make the very controversial assumption that our clear, intuitive content-attributions constitute strong evidence for subjects' having mental representations with the attributed contents, our intuitive judgments still provide little help for developing stability constraints--we *just don't seem to have any clear ones* about cases that would yield useful information for developing stability constraints.

To illustrate this, consider the following case. Say that there's a version of yourself, called 'Self$_0$', that is psychologically identical to you except that Self$_0$ believes that jazz keyboardist and composer Sun Ra's album *Lanquidity* was released in 1978. Conceive of a series of Selves such that Self$_{n+1}$ is psychologically identical to Self$_\mathrm{n}$ but with one of her beliefs "erased" from her psychology.[15] Now, say that Self$_k$, where k is some relatively large number, has lost so many beliefs that she couldn't tell you what an album is or what jazz is or that 1978 came after 1975. According to our folk content-attributions, Self$_k$ does not believe that *Lanquidity* was released in 1978. We assumed that Self$_0$ does believe that *Lanquidity* was released in 1978, so somewhere in the series from Self$_0$ to Self$_k$ there is some Self that is the first in the series to lack this belief. According to your intuitions about content-attributions, where is this Self? At what point does Self stop believing that *Lanquidity* was released in 1978? For what n would Self$_n$ no longer be able to agree with Self$_0$?[16]

It seems, at least to Stich (1983 pp. 85-6) and me, that our pre-theoretical intuitions about content-attributions provide us with little, if any, guidance about how to answer these questions. But notice that these types of questions are exactly those that are relevant for developing useful

[14] For instance, if our folk content-attributions say that genuine disagreement can *always* occur between subjects with representational systems with property K (e.g., the property of sharing most but not all beliefs about cats) and theory of content T implies that systems with property K can only *rarely* disagree, then we have reason to reject T.

[15] Assume that beliefs are erased according to some "centrality" ordering in relation to Self$_0$'s belief that *Lanquidity* was released in 1978; the most peripheral are erased first, etc.

[16] Cf. Stich's (1983 pp. 55-6, 85-6) case of Mrs. T.

stability constraints. Our folk content-attributions are supposed to provide evidence for when and how often disagreement, changes of mind, etc. actually occur. But the above suggests that, beyond the extremes (e.g., Self_0 can't agree with Self_k), our folk content-attributions provide no clear answer in a given case (e.g., If Self_{450} doesn't have any beliefs about keyboards can she agree with Self_0? I can only shrug).[17]

To sum up this chapter, two lines of thought suggest that we should appeal to the explanatory roles of representations in developing stability constraints. First, we have a clear and compelling *motivation* to do so because of the explanatory success of cog sci theories that posit representations. Second, we seem to have *no promising alternatives*; we have reason to believe that our pre-theoretic, folk content-attributions will be of little help for developing theoretically helpful stability constraints. In the next chapter, I will say more about *how* we can appeal to the explanatory roles of representations and content in cog sci theories to develop useful stability constraints for theories of content.

[17] At the very least, the burden is on the advocate of folk content-attributions to provide some reason why, in light of Self-like cases, we should not be so pessimistic about developing useful stability constraints based on folk content-attributions.

Chapter 4: A rough procedure for extracting stability constraints from cognitive science theories

For the sake of simplicity, in what follows I will restrict my focus to cog sci theories that qualify as *classical computational* theories of cognition. Roughly, these theories hold that representations occur at the symbol-level and have combinatorial, "language-like" syntactic and semantic structures (see Fodor & Pylyshyn 1988). Moreover, these theories hold that cognition is computation; cognitive systems perform tasks by manipulating simple and complex symbols so conceived. This restriction to classical computational theories simplifies things because, generally, it is less controversial that contentful representations figure in these theories as genuine explanatory posits. It is more controversial whether, say, connectionist theories posit representations (see Ramsey 1997; 2007).[18] So, hereafter I will refer to classical computational theories of cognition by 'cog sci theories'.

In Chapter 3, I said that a theory of content is adequate only if it is compatible with the explanatory role of representations in successful cog sci theories. Put differently, we should reject a theory of content that implies that an independently established, successful cog sci theory is unsuccessful. There are certain necessary conditions that a cog sci theory must meet to be explanatorily successful, so a theory of content that would undermine an established cog sci theory's ability to meet these conditions is ipso facto inadequate. One way in which a theory of content might do this is by implying that when representations are manipulated by formal procedures specified by the cog sci theory, their contents change so that the cog sci theory is no longer explanatory. This suggests the following rough procedure for formulating stability constraints:

(1) Find a well-established, explanatorily successful cog sci theory, in which contentful representations figure.

(2) Determine the explanatory role that content plays in this theory.

[18] But this does not mean that what I say below is inapplicable to non-classical-computational theories. Insofar as these theories do invoke representations as genuine explanatory posits, my comments apply *mutatis mutandis*.

(3) This explanatory role induces a set of necessary conditions *C*, which must be met in order for content to play this role and for the cog sci theory to be explanatorily successful. Determine *C*.

(4) Take stock of the formal procedures performed on representations as specified by the cog sci theory.[19]

(5) Formulate *Stability Constraint*: If a theory of content *T* entails that any of the formal procedures performed on representations specified by the cog sci theory are *content-changing* such that some member of *C* is no longer met, then *T* is inadequate.

To illustrate how this procedure should work, we will look at two kinds of explanation, which some representation-positing cog sci theories are instances of: task-decompositional and model/simulation explanations. I will focus on the different explanatory roles that content plays in these explanations and the necessary conditions that these explanatory roles induce (i.e., *C* in (3) and (5)). Unfortunately, space limitations forbid detailed investigations of particular cog sci theories, so my comments will remain at a somewhat general level focusing on *kinds* of explanations rather than particular instances of these kinds. Hopefully it will be clear how the rough procedure applies to particular task-decompositional and model/simulation cog sci theories.

4.1 The set-up: the anti-representationalist challenge

Following Egan (2012 p. 254), cog sci theories that posit representations construe the mind as an information-using device. Here, information-using devices are understood as *physical symbol systems* (PSS), devices that manipulate symbols according to instructions provided by a program. Symbols are *physically realized* objects with *semantic content*; accordingly, for a given (type of)

[19] The characterizations of these formal procedures will need to include more than just the bare formal properties of the representations that they apply to so that the resultant stability constraints do not turn out to be too exclusive. What should be included in these characterizations will depend in part on the theory that the procedures are taken from, but what I have in mind is something like the "task context" in which the procedures are normally deployed according to the theory. By 'task context', I mean something like the task (or cognitive capacity) that the formal procedure is supposed to be contributing to by being performed on the relevant representations. The situation that I want to rule out is one where, for example, a model of our language parsing capacities has some procedure *p* that applies to representations that have similar formal properties to those that a formal procedure *q* applies to, but *q* is embedded in a model for our decision-making capacities. It would be inappropriate to reject a theory of content *T* just because *T* implies that representations with contents relevant to decision-making would problematically change if the language-parsing procedure *p* were applied to them.

PSS, there is a realization function f_R, which maps symbols to physical state-types of the system and an interpretation function f_I, which maps symbols to their semantic contents (objects, properties, propositions, etc.). According to this notion of a PSS, a *computation* "is a sequence of physical state transitions that, under the mappings f_R and f_I, executes some specified task" and a *representation* "is an object whose formal and semantic properties are specified by f_R and f_I respectively" (Egan 2012 p. 254).

Cog sci theories explain intelligent behavior and capacities in terms of procedures performed on (physically realized) symbolic representations. But, importantly, these procedures apply to representations *only* in virtue of their *formal* properties, not their semantic properties (see Fodor 1980). The procedures performed on representations are defined without reference to those representations' contents; they are defined only with reference to representations' physically realized, formal properties. Formal properties "encode" semantic properties, that is, "formal features mirror (in the sense of bearing a one-to-one correspondence with) semantic characteristics of some represented domain" (Pylyshyn 1980 p. 114). But, seemingly, all that is required for a *sufficient causal explanation* of some behavior is the relevant formally defined algorithm, the values of $f_R(x)$ for all symbols implicated by the algorithm, and perhaps functional characterizations of those values of $f_R(x)$ in accordance with the formal relations between the symbols. We need not advert to any contents in specifying a causal explanation for some behavior--all of the "causal work" is done by the physically realized, formal properties of representations. Consequently, as Fodor (1982) puts it,

> ...it looks as though [the notion of the content of a mental representation] is never going to be required in order to give the explanations that cognitive scientists want to give. The idea that mental operations are formal can thus be taken to imply that the content of a mental representation is a dispensable construct, at least for the purpose of cognitive science (p. 100).

This provides a challenge to those who think that representations with content play a genuine role in explaining intelligent behavior and capacities. If, apparently, we can give full causal explanations of behavior without appealing to representations' semantic properties, what is the motivation for saying that they have them? What explanatory purposes do the contents of representations serve that can't be served by representations' formal properties alone?

4.2 The role of content in task-decompositional explanations[20]

Generally, the explananda of cog sci theories are cognitive capacities, e.g., deductive reasoning, understanding analogies, language learning, etc. Cognitive capacities are typically characterized as "input-output conversion[s] couched in representational terms" (Ramsey 2007 p. 69). Thus, in order for the computational models posited by cog sci theories to provide adequate explanations of these explananda, their inputs and outputs must be interpreted in the right way. The computational model for deductive reasoning must have inputs that are interpreted as *premises* and outputs interpreted as *conclusions*; the model for multiplication must have inputs interpreted as *multiplicands* and outputs interpreted as *products* (Ramsey 1997 p. 38). The syntactically described representations (i.e., data structures) that constitute the model's inputs and outputs need to be interpreted as *representing*, e.g., numbers, in order for the model to be an explanation of the target explanandum, e.g., the capacity to do multiplication. Otherwise, the computational device posited by the theory does not perform *multiplication*, and thus, does not explain a cognitive system's capacity to do multiplication.

However, this alone is not sufficient to meet the anti-representationalist challenge. The characterization of a cognitive capacity as a conversion of representational inputs and outputs is a *pre-theoretical* characterization of the *explanandum* of a cog sci theory. It is not part of the theory proper, the *explanans* (Ramsey 2007 p. 71). Consequently, we have not yet articulated a sufficient motivation for treating the structures posited by a theory as contentful representations. The anti-representationalist could say that the explanatory work done by the theory proper is accomplished without appealing to any semantic properties of the theory's posited structures. We can treat the overall, "external" inputs and outputs of the computational model as representational, but, says the anti-representationalist, this is only an *informal* way of motivating the theory by connecting it to our pre-theoretical conception of the cognitive capacity in question; it is not part of the explanatory apparatus of the theory proper (cf. Chomsky 1995). To answer the anti-representationalist challenge, we need a reason for treating some of the *internal* structures posited by the model as contentful representations.

According to some (e.g., Cummins 1975; 1983), computational models explain cognitive capacities by showing how the complex system that performs the relevant task (or possesses the

[20] In this subsection and in the next, I will take most of the exposition of the views on the explanatory role of content from Ramsey (2007 Ch. 3).

relevant capacity) is composed of less sophisticated sub-systems whose organized functioning produce that complex task (or allow the complex system to have that capacity) (Ramsey 2007 p. 71).[21] For example, the system described by the computational model for multiplication might be composed of sub-systems that perform repeated addition. But importantly, just as it is necessary to interpret the inputs and outputs of the overall system posited by the multiplication model as representing numbers in order for the system to be performing *multiplication*, in order for its sub-systems to perform *addition*, it must be the case that *their* inputs and outputs represent numbers, specifically, addends and sums (ibid. p. 72). Otherwise, we "won't be able to view the sub-system as an adder, and hence we won't be able to see how and why its implementation is essential to the overall capacity being explained" (ibid.). Task-decompositional explanations in cog sci work by showing how less sophisticated sub-systems' organized functioning contribute to a super-system's cognitive capacity, that is, an ability to convert representational inputs and outputs in the right way. But, these explanations are explanations of these *cognitive* capacities only if the sub-systems are appropriately characterized as *themselves* converting representational inputs and outputs in the right way; otherwise it's not clear how these sub-systems' functioning contribute to the representationally characterized cognitive capacities. So, the explanatory role of contentful representations posited throughout the computations and sub-computations of a task-decompositional cog sci theory is to allow that theory to explain *cognitive* capacities, i.e., capacities characterized as processing or transforming *representations*.

In order for representations to play this role, their content-assignments must meet certain stability requirements. To see this, say we have a PSS that performs multiplication and is contained in a black box; when given inputs that, under an interpretation, are certain numbers, it outputs the correct products (cf. Haugeland 1978). A task-decompositional explanation of this PSS's capacity to perform multiplication would provide a description of the organized functioning of its sub-systems (i.e., an algorithm) that allows it to produce the right outputs given arbitrary inputs.[22] And, the various sub-systems cited in the explanation will be described as performing various procedures on representations with contents.

However, if a theory of content *T* entails that any of the various procedures performed by the sub-systems are *content-changing*, such that the resultant output of the black-box-PSS is a

[21] Compare to *constitutive explanations*, which explain the capacity of a mechanism, X, by describing the organized function of X's component parts (Craver 2001 p. 70; Salmon 1984).

[22] Or at least, inputs included in the set of cases required to explain the capacity in question.

representation with the *wrong* content in some case required to explain the PSS's capacity to perform multiplication, then *T* undermines the explanation and thus, is implausible. In order for the proposed task-decompositional explanation to be successful, its algorithm must produce the correct outputs given arbitrary inputs; for example, according to the multiplication algorithm it must be the case that (3, 3) → (9). If *T* entails that one or more representations changed their contents as a result of one or more procedures in the black box such that the content of '9' is, say, *8*, then *T* entails that the proposed explanation is inadequate because it produces incorrect outputs. According to *T*, representations do not meet the stability requirements for fulfilling their role in this task-decompositional explanation.

4.3 The role of content in model/simulation explanations

How does a map help us navigate and make inferences about a landscape? It has symbols, lines, shapes, etc. that bear relations to one another in such a way so that they mirror the relations that objects in the landscape bear to one another; there is a structural isomorphism between the symbols on the map and the objects in the landscape. This insight, viz., structural symmetries between a map and a landscape can be exploited to produce knowledge and perform tasks, is employed in explanations of cognitive capacities. Theories posit representations--symbol structures--that constitute models (i.e., simulations, surrogates) for something in the real world--target domains. These representations constitute models or simulations of their target domains in virtue of being structurally isomorphic to those target domains; the relations between the symbols that compose the models mirror the relations between objects in the target domain. Cognitive capacities are explained in terms of formal procedures performed on the constituent symbols of the models that allow cognitive systems to exploit these structural symmetries. The "surrogative reasoning" performed on the models in the form of formal procedures allows cognitive systems to make inferences, perform tasks, solve problems, etc. in/about the target domain (Ramsey 2007 pp. 78-83).

So, why must we construe the symbols that make up these models as representations of things in the target domain? Why can't these explanations of cognitive capacities be composed only of the formal procedures performed on symbols? Cog sci theories set out to explain *how* a cognitive capacity is implemented in a system, but they also set out to explain *why* systems are able to perform the tasks they do. The latter is left utterly mysterious without the assumption that the symbols of models are *representations of* things in the world. A purely syntactic account doesn't answer the question "why do *these* syntactic operations enable the system to perform as well as it does?" (ibid. p. 90) So, the explanatory role of representations with content in model/simulation cog sci explanations is to provide an answer to this question: Why are cognitive systems successful at performing certain tasks?

If we focus on this question, the stability requirements induced by this explanatory role of representations fall out quite nicely. What are the properties of models/simulations posited by cog sci theories that allow cognitive systems to perform tasks successfully? Just as with maps of landscapes, models are useful only if their symbols are structurally isomorphic to their target domains.[23] Cog sci theories can answer the question of why cognitive systems are successful at performing tasks only if the models they posit are *in fact conducive to systems performing tasks successfully*. And models are in fact conducive to systems performing tasks successfully only if their symbols are structurally isomorphic to what they represent.

Thus, if according to a theory of content *T* a formal procedure on symbols of a model specified by an otherwise successful cog sci theory is *content-changing* such that the internal structural relations between the symbols of the model are no longer isomorphic to the target domain in some case required by the explanandum, then the representations posited by this theory do not meet the stability requirements. Representations with contents determined according to *T* are inadequate for their explanatory role; they are not isomorphic to their target domain over the formal procedures specified by the theory in the cases required to explain the capacity in question. And consequently, the cog sci theory cannot provide an adequate answer to the question of why cognitive systems are successful at performing the task(s) that the theory sets out to explain. If representations' contents are determined according to *T*, then the model lacks the property (viz., being structurally isomorphic to the target domain) that allows cognitive systems to exploit it to perform the task(s) successfully.

[23] On some readings of 'useful', this conditional is false; whether a map that is not isomorphic to the target domain is useful depends (in part) on the purpose(s) that it is used for and the extent to which the map deviates from being isomorphic to the target domain. I can still use my slightly distorted map of the state of Virginia to get around the western part of the state even if Richmond is shifted to the north so that the map is no longer isomorphic to the state. Similarly, whether a model posited by a cog sci theory is useful for the cognitive system will depend on the task and the theoretical context. So, the stability requirements induced by the explanatory role of content are a bit messier than I let on here. However, I do not see this as an urgent problem. Even if strict isomorphism is not a necessary condition for usefulness and thus, for the explanatory success of the theory, there will likely be other, similar necessary conditions that will be fixed by the theory, its explanandum, etc.

4.4 A quick summary of the information in this chapter relevant to performing the rough procedure

Task-decompositional explanations

Explanatory role of content: allows computational models to explain *cognitive* capacities

Relevant necessary condition:[24] the outputs of the model must have the correct contents in cases required to explain the capacity in question

Stability constraint guideline: If a theory of content *T* entails that any of the formal procedures performed on representations specified by the task-decompositional cog sci theory are *content-changing* such that the contents of the outputs of the computational model posited by that theory are incorrect in some case required to explain the capacity in question, then *T* is inadequate.

Model/simulation explanations

Explanatory role of content: allows theories to explain *why* cognitive systems are able to perform the relevant tasks

Relevant necessary condition: the model posited by the theory must be structurally isomorphic to the target domain in cases required to explain the capacity in question

Stability constraint guideline: If a theory of content *T* entails that any of the formal procedures performed on representations specified by the model/simulation cog sci theory are *content-changing* such that the model posited by the theory is no longer structurally isomorphic to the target domain in some case required to explain the capacity in question, then *T* is inadequate.

[24] In other words, the member of the set *C* in the rough procedure (the set of necessary conditions for content to play its explanatory role) that appears to be at risk of being undermined by unstable contents. In principle, if any member of *C* is not met according to a particular theory of content, then we have reason to reject that theory. However, for the sake of the feasibility of the rough procedure, it is not required that we be able to list all members of *C* in the resultant stability constraint; we only need to identify those that are at risk of being undermined by unstable contents.

Chapter 5: Conclusion

In this paper, I have done three things that should contribute to developing a holistic theory of content that does not entail content instability. i) I argued that the strategy suggested by the Many-One response, viz., appealing to properties of the determination functions of theories, is a non-starter for developing useful stability constraints. ii) I proposed a more attractive alternative, viz., adverting to the explanatory role(s) of representations in cog sci theories. iii) I outlined and illustrated a rough procedure for extracting useful stability constraints from cog sci theories.

I think that it is too early to say whether we should be optimistic about the prospects for developing a plausible holistic theory of content. We need a clearer view of the stability constraints coming out of particular cog sci theories. I have outlined how, following the rough procedure, such constraints can be generated from theories that qualify as task-decompositional or model/simulation explanations. Once we generate these constraints, we can evaluate otherwise plausible holistic theories of content in terms of them. Through this process, we can get a more detailed picture of just how problematic the problem of content instability really is for holistic theories and of which types of theories (if any) have the resources to surmount it.

Appendix A: Showing that many-one determination functions are insufficient for avoiding content instability

In this Appendix, I will illustrate the problem with the many-one stability criterion that I identified in Chapter 2 with an example. I will show that a theory's having a many-one determination function is insufficient for avoiding content instability. As I go through this example, it will help to keep in mind the "procedure" I sketched for turning a theory with a one-one determination function into a theory with a many-one determination function that entails the same degree of content stability. Roughly, I will be "smuggling" the one-one theory's determination base individuating conditions into another theory's many-one determination function whose determination bases are much more finely individuated.

First, we need the seemingly uncontroversial premise that two theories T_1 and T_2 entail the same degree of stability if T_1 and T_2 get the same stability results. That is, every case in which a system is exposed to a perturbation and its content assignments change in a certain way according to T_1 is a case in which that same system, exposed to the same perturbation, changes its content assignments in that same way according to T_2. And the same thing can be said for cases in which a system is exposed to a perturbation and its content assignments do not change according to these theories.

Now, suppose that according to holistic theory *H**, the content of a representation *r* for a system *Q* is (wholly) determined by *r*'s inferential role in *Q*. And according to *H**, the inferential role for a representation *r* is the set of all inferences in which *r* figures where the conditional probability of the conclusion being tokened given the tokening of the premises is greater than *n*. To make this a bit more tractable, we can assume that *Q* is a mental representational system, say, a subject *S*'s representational system, and we will say that the inferential role of a representation for *S* according to *H** is the set of inferences that *S* is disposed to make. To simplify for illustrative purposes, we will say that *S* is disposed to make an inference (according to *H**) from *A* to *B* iff $P(B|A) > n$. That is, *S* is disposed to infer *B* from *A* iff the probability of *S* tokening *B* given a tokening of *A* is greater than *n*.

For *H**, the determination function *h** is one-one. Consequently, *H** entails **Change**: any change in the inferential role of *r* (i.e., any change in the inferences *S* is disposed to make with *r*) changes the content of *r* and any representations *r* is related to in its inferential role. Thus, *H** entails content instability.[25]

Now, suppose that there is another holistic theory *H*** that defines the determination base for the content of *r* as the set of all *S*'s inference *tokens* in which *r* figures. Furthermore, *H*** defines a determination function *h*** that assigns the same content to two representations r_1 and r_2 iff the set of all r_1-inference-types[26] with $P(B|A) > n$[27] is identical to the same such set for r_2. In other words, $h^{**}(r_1)$ and $h^{**}(r_2)$ take as inputs the total sets of inference tokens for r_1 and r_2, respectively, and the outputs of $h^{**}(r_1)$ and $h^{**}(r_2)$ are the same iff the set of all r_1-inference-types with $P(B|A) > n$ is the same as the set of all r_2-inference-types with $P(B|A) > n$.

Since the domain of *h*** is all the possible sets of *S*'s inference tokens involving *r* and, multiple, distinct sets of inference tokens can determine the same set of inference-types with $P(B|A) > n$[28] (for all such sets of inference-types) *h*** is strongly many-one. *H*** does not entail **Change**: there are changes in *r*'s determination base that do not change *r*'s content. For example, an additional tokening of an *r*-inference-type that does not lower the conditional probability of any *r*-inference-type that was $> n$ such that its conditional probability is now $\leq n$ or raise the conditional probability of any *r*-inference-type that was $\leq n$ such that it is now $> n$.

However, if *H** entails content instability, so does *H***, for every change in *S*'s inferential role for *r* according to *H** (and consequently a change in content) would be a change in content according to *H***. The determination function *h** assigns the same content to representations iff they have the same inferential role, i.e., the same set of inference-types for which $P(B|A) > n$, and *h*** assigns the same content to representations iff they have the same set of inference-types for which $P(B|A) > n$. So, since *H** and *H*** entail the same content changes in all the same cases and they entail content changes in no other cases, by our "seemingly uncontroversial premise" above, they entail the same degree of content stability. Thus, since *H* does* entail content instability, *H*** does too. *H*** is a theory with a strongly many-one determination function, so a

[25] To run a successful instability objection against *H** we would need to support the relevant premise (ii). This would require the assumption that for most representations, virtually no two people share all the same inferential roles; i.e., inferential dispositions (as defined by *H**). This claim seems pretty uncontroversial, assuming that *n* isn't too high (say, greater than .9). For the sake of argument, we can assume that *n* is such that we could justify the relevant premise (ii) in the instability objection against *H**.

[26] As determined by the set of all *S*'s r_1-inference tokens.

[27] *A* is the set of premises, and *B* the conclusion of an inference-type in which r_1 figures.

[28] Here's a simple example: Say $n = .4$ and we have two sets of inference tokens {Q→R, Q→T} and {Q→R, Q→R, Q→T, Q→T}. For both these sets of tokens $P(\text{R}|\text{Q}) = .5$ and $P(\text{T}|\text{Q}) = .5$, so they have the same set of inference types whose $P(B|A) > .4$, viz. {Q→R, Q→T}.

theory's having a strongly many-one determination function is not sufficient for avoiding content instability.

References

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy* , 615-678.

Block, N. (1995). An Argument for Holism. *Proceedings of the Aristotelian Society* , 151-169.

Block, N. (1993). Holism, Hyper-Analyticity and Hyper-Compositionality. *Philosophical Issues* , 37-72.

Chomsky, N. (1995). Language and nature. *Mind* , 1-61.

Churchland, P. M. (1998). Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. *The Journal of Philosophy* , 5-32.

Churchland, P. M. (1986). Some Reductive Strategies in Cognitive Neurobiology. *Mind* , 279-309.

Churchland, P. M. (1993). State-Space Semantics and Meaning Holism. *Philosophy and Phenomenological Research* , 667-672.

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science* , 53-74.

Cummins, R. (1975). Functional Analysis. *Journal of Philosophy* , 741-60.

Cummins, R. (1983). *The Nature of Psychological Explanation.* Cambridge, Mass.: MIT Press.

Dresner, E. (2012). Meaning Holism. *Philosophy Compass* , 611-619.

Egan, F. (2012). Representationalism. In E. Margolis, R. Samuels, & S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 250-72). Oxford University Press.

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

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. In S. Pinker, & J. Mehler (Eds.), *Connections and Symbols.* Cambridge, Massachusetts: MIT Press.

Fodor, J. (1982). Cognitive Science and the Twin-Earth Problem. *Notre Dame Journal of Formal Logic* , 98-118.

Fodor, J. (1980). Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioral and Brain Sciences* , 63-73.

Fodor, J., & Lepore, E. (1999). All at Sea in Semantic Space: Churchland on Meaning Similarity. *The Journal of Philosophy* , 381-403.

Fodor, J., & Lepore, E. (1992). *Holism: A Shopper's Guide.* Cambridge: Blackwell.

Fodor, J., & Lepore, E. (1993). Reply to Critics. *Philosophy and Phenomenological Research* , 673-682.

Harman, G. (1987). (Nonsolipsistic) Conceptual Role Semantics. In E. Lepore (Ed.), *New Directions in Semantics.*

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences 2* , 215-260.

Jackman, H. (1999). Moderate Holism and the Instability Thesis. *American Philosophical Quarterly* , 361-369.

Lormand, E. (1996). How to Be a Meaning Holist. *The Journal of Philosophy* , 51-73.

McLaughlin, B. P. (1993). On Punctate Content and on Conceptual Role. *Philosophy and Phenomenological Research* , 653-660.

Pagin, P. (1997). Is Compositionality Compatible with Holism? *Mind and Language* , 11-33.

Pagin, P. (2006). Meaning Holism. In E. Lepore, & B. C. Smith (Eds.), *The Oxford Handbook of Philosophy of Language* (pp. 213-232). New York: Oxford University Press.

Pitt, D. (2012). *Mental Representation*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/entries/mental-representation/

Pylyshyn, Z. (1980). Computation and cognition: issues in the foundations of cognitive science. *The Behavioral and Brain Sciences* , 111-169.

Ramsey, W. (1997). Do Connectionist Representations Earn Their Explanatory Keep? *Mind and Language* , 34-66.

Ramsey, W. M. (2007). *Representation Reconsidered.* Cambridge: Cambridge University Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Stich, S. (1983). *From Folk Psychology to Cognitive Science.* Cambridge: MIT Press.

Tiffany, E. (1999). Semantics San Diego Style. *The Journal of Philosophy* , 416-429.