Teleosemantics and Swampman: Defanging an intuition[1]

1. Introduction

Teleosemantics is one of the leading attempts to naturalize intentionality. In an informal survey of my naturalistic colleagues, I found that the most important factor preventing them from adopting teleosemantics is the infamous Swampman problem. Suppose a molecular duplicate of Donald Davidson, against all probability, self-assembled in the Florida Everglades after a lightning strike. Assuming teleology is endowed by a history of evolution and learning, Swampman's mental states lack teleology. Therefore, according to teleosemantics, either his mental states lack content, or he has no mental states at all (except perhaps phenomenal states). My colleagues feel that this is so wildly implausible that it amounts to a *reductio* of teleosemantics.

In this paper, I will try to defang the Swampman intuition. There are two extant approaches to doing so. The first is to argue that teleosemantics is an empirical theory of the nature of representation, along the lines of the empirical theory of water as H_2O (Papineau, 2001). Thus we should disregard the intuition that Swampman has truth evaluable mental states just as we should disregard the intuition that twin-water (XYZ) is water. While I think this argument is fundamentally correct, it is strengthened in combination with the second strategy, as exemplified by Fred Dretske's example of Twin-Tercel (Dretske, 1996). Twin-Tercel is a whirl-wind induced spontaneous assembly of a car molecularly identical to Dretske's own Toyota. But its gas gauge lacks the function of indicating the level of gas in the tank, so it doesn't represent the level of gas

a role in properly truth-evaluable contents, similarly require causal contact with their denotations, it will be more plausible that Swampman's mental states are devoid of content. The difficulty is to motivate such a strong, global externalism.

The second prong is to show that, although Swampman's mental states lack truth evaluable content, they do possess something analogous which allows us to talk about Swampman's behaviour in much the same way that we would speak about Davidson's. This "content" by *fiat* is a pragmatic sort of "content" founded upon the isomorphism of Swampman's mental states to his environment (an isomorphism shared by Davidson's mental states). The difficulty here is to demonstrate the isomorphism.

The third prong is to demonstrate that Swampman *has mental states*, and not only purely phenomenal ones. The final burden of the paper, then, is to show how it is possible to have perceptions, beliefs, and desires that lack truth evaluable content. Demonstrating this will rest on an independently motivated psychofunctionalism as applied to any representational theory of mind.

The three prongs rest on independent motivations for global externalism, isomorphism between mental representations and the world, and psychofunctionalism. Just as Dretske's Twin-Tercel argument rests on an independent motivation for his indication-based teleosemantics, my three prongs will be justified by an independently motivated isomorphism-based teleosemantics. This motivation comes from psychology (as it did for Dretske, especially the psychology of reinforcement learning), but more importantly from neuroscience. As this story is told & justified in detail elsewhere [references omitted for blind review], and time is limited, I'll just hit the necessary highlights as we proceed.

2. Model representation, model building machines, and SINBAD

Ordinary artifact models represent by normative isomorphism. For example, a model airplane represents the kind of plane it does because it is *supposed* to be spatially isomorphic to it; a rock

Similarly in Figure 1[2], if the small hat represents the big hat, it does so not merely because its spatial structure *actually* mirrors the spatial structure of the big hat, but because its spatial structure is *supposed* to mirror the spatial structure of the big hat. More generally, it represents the big hat because it is normatively isomorphic it.

According to the SINBAD theory, the brain builds such isomorphisms in a way analogous to how the machine in Figure 1 does - the automatic scale modeler. This machine takes an object through its input door, makes a mould of the object, shrinks the mould, injects a fast-hardening plastic, and voilà! there you have a scale model. There are two things that determine the representational content of a particular model this machine spits out. First, the template object that causes production of the model is relevant, i.e. the history of model. In Figure 1, the model is a model of that hat because that hat was its template. Second, the design principles of the machine are relevant. The automatic scale modeler is not designed to mirror colour structure, only spatial structure. For instance, the model hat in figure 1 doesn't misrepresent the colour of the template hat, because the machine isn't *supposed* to produce colour isomorphisms. The template object and design principles together determine that the model in Figure 1 represents the shape of that brown hat.

Figure 1: The automatic scale modeler

[2] Actually a presentation slide from the talk.

According to the SINBAD theory, the brain is a model-building machine, but it's designed to produce isomorphisms, not to spatial structures, but to regularity structures - like an orrery, which is a dynamic model, isomorphic to solar system regularities. This isomorphism is useful for making predictions by what's usually called "filling in". For example, if you know where Earth will be in two months, but not Venus, you can *fill in* this missing information by rotating Earth into its known position. The gears of the orrery will allow you to "read off" the future position of Venus.

If we want our model building machine to build dynamic models, we need environmental regularities to be templates for the production of internal, mirroring regularities - as in a classical associationist system. However, the cerebral cortex appears to be designed to mirror a more complex regularity pattern than pairwise correlation, namely a clustering pattern.

Regularities tend to cluster, as described by Boyd, Kornblith, and Millikan in their related accounts of natural kinds: the unified property cluster account. Normally, the clustering occurs for an underlying reason: for example the properties of water cluster together due to its chemical structure, and the properties of cats cluster together due to individual cats sharing an evolutionary history. In information theory terms, the properties of water and of cats are "mutually informative", so I call these kinds "sources of mutual information" (or SOMIs). Following Millikan, we can extend the notion of a source of mutual information to include non-natural (but nevertheless real) kinds as well. The properties of screwdrivers cluster in part because they serve a specific function; the properties of Powerbooks cluster because they originate from the same plan. Millikan has shown that even individuals fit the pattern.

environment, allowing for the all-important "filling in" process. The dynamic isomorphism that develops mirrors the *deep structure* of the environment, with elements that correspond to the individuals and kinds - that is, the sources of mutual information - around which environmental regularities are organized.

I have argued elsewhere [references omitted for blind review] that this isomorphism is also normative[3] - SINBAD networks *represent* the deep regularity structures of the environment, and the SOMIs around which they're organized. The cortex is a dynamic model-building machine that is designed by evolution to produce models of regularities involving sources of mutual information. The main design principles of this machine are given by the SINBAD theory, allowing us to identify the templates for particular models the cortex produces. Just as a product of the automatic scale modeler represents the spatial structure of its template object, a portion of the SINBAD cortex represents the template regularity structure that has been imprinted on it.

3. The first prong: global externalism

We are now in a position to appreciate the first prong of my argument designed to attenuate the intuitive pull of the Swampman case: an independently motivated global externalism. According to the SINBAD theory, the entire cortex is structured through the construction of isomorphisms through environmental regularities "imprinting" themselves into a specially designed medium. This applies equally to early visual cortex, where cells tune to discontinuities in reflected & emitted light (Jones and Palmer, 1987), as it does to inferotemporal cortex, where cells apparently tune to object kinds

to the cell group that mimics Davidson's "Mom" cells as it does to the group that mimics Davidson's "red"[5] cells.[6]

There are a few places one could press. Most obviously, my contention is hostage to the empirical appropriateness of the idea that cortical learning is the template-based production of isomorphisms. Fair enough; but SINBAD is backed up by considerable empirical support, and further, it is only one of many possible such models. Second, supposing that the SINBAD theory is correct, the analysis of model representation as normative isomorphism could be wrong. (I doubt it.[7]) Third, perhaps the isomorphisms developed in cortical SINBAD networks are not normative, despite my arguments to the contrary [reference omitted]. Fourth, while the cortex could be populated by SINBAD model representations, the relation between these representations and mental representation could be more complex than the simple identity canvassed here. (In this connection, I point you to the explanatory power of the SINBAD idea with respect to folk psychology [reference omitted], some of which we'll see in a moment.) Relatedly, perhaps some mental representations (e.g. of cause, of necessity, of number) outstrip SINBAD's capabilities, and must receive some other explanation. (One could respond by narrowing the scope of teleosemantics, although I do not think this is necessary.) However, IF the theory avoids those problems (as I think it does), then the global externalism that follows should substantially weaken the Swampman intuition against this version of teleosemantics.

4. The second prong: isomorphism

Davidson presumably had a pretty accurate internal model of his house. When he made use of this model (more on how in the next section), he could engage in what an outside observer would consider to be successful behaviour with respect to his house. The same, of course, applies to Swampman. The difference is that Davidson's internal model is *supposed* to apply to the house, whereas Swampman's structure, although it exhibits a high degree of isomorphism with

---

[5] With di!erent optics and/or di!erent photoreceptive equipment, a SINBAD network would develop a very di!erent set of colour representations.

[6] I note that this is all compatible with there being a nativist element to perceptual and conceptual development. Evolution can of course add

Davidson's house, isn't supposed to be isomorphic to anything at all. Just as I can (stupidly) use a rock that I happen to find while hiking to fill in missing information about the spatial structure of Davidson's house, Swampman can use his isomorphic structure to fill in missing information about it. But neither the rock nor Swampman's isomorphic structure represent Davidson's house – the accuracy of the application is a fluke in both cases.

However, flukish accuracy is still accuracy, and this can ground a pragmatic sort of content, to be

of Swampman.[8] Notice, also, that if Davidson were whisked away to some planet somewhat similar to Earth, we could easily determine which of his perceptual judgements were false, but not so with Swampman. In fact, there is no reason to say that Swampman has an accurate model of

hot and which cold, the temperature at the tap, and the flow rate at the tap. These variables covary in regular ways as governed by the causal structure characteristic of sinks. For instance, if I turn the left knob by a certain number of degrees, the temperature at the tap will change by a corresponding amount in a direction dependent upon whether the left knob is hot or cold. A dynamic model of the sink will have elements that represent or "stand in" f "st3

The action-guiding functional mode is a bit more complicated. In building its internal model, the organism's internal modeller learns not only about regularities in the outside world, but also about how variables in the environment relate to its own needs and satisfaction. It learns, for instance, that when it has need N, and the environment is in (complex) condition W, satisfaction is high. When we use an *external* model of the sink, we must "tell" the model what tap variable values we want, and read off what the knob positions need to be. In the internal model of an autonomously acting organism, instead of desired tap variable states, a high value of "satisfaction" is the (sham) input. The internal model will then fill in, given the organism's current needs (e.g. basic drive signals)ic drive

in dreaming - when the network operates free from both input and behavioural output; that is, an exploratory mode corresponding to the attitude of supposition.)

Therefore the SINBAD theory is a perfect fit for the standard RTM[11] account of the attitudes: the same representation occupies different causal roles in order to implement beliefs, desires, etc. Importantly, these roles must characterize attitude types independent of their semantics. (This is especially obvious for the many versions of RTM - most notably Fodor's - that reject a causal role semantics.) The SINBAD model gives us a way of understanding how this works, where thalamic switching mechanisms control the flow of information within the system, determining whether the network region in question is operating in indication (judgement) or direction (occurrent desire) mode. (Dispositional beliefs, and desires proper can be characterized in relation to judgements and occurrent desires - they are judgements the system is disposed to make, or occurrent desires the system is disposed to implement, given certain inputs.[12])

More remains to be said about the large subject of the propositional attitudes (for some of this, see [reference omitted for blind review]). But perhaps this is enough to see how Swampman could make a judgement that lacked any truth evaluable content. If the modes of use are characterized, not teleologically, but as causal roles (a psychofunctionalism that is standard in RTM), there is no reason why Swampman's isomorphic internal structures cannot occupy the causal roles characteristic of beliefs and desires. They are not like Davidson's beliefs and desires, because their constituting elements (analogs to Davidson's concepts) have no templates, and so have no determinate contents (section 3 above). But they do exhibit isomorphisms to the environment, and so may be assigned *pragmatic* contents (section 4). (Assuming Davidson was not greatly deceived about things, these will largely match the corresponding properly truth-evaluable contents that characterized Davidson's attitudes.) And they occupy the characteristic causal roles of judgement and occurrent desire (or dispositional belief and desire proper) courtesy of the bare causal structure of Swampman's brain analog. What more do you want?

[11] Representational Theory of Mind. structnd.

[12]

All of this rests upon whether or not the SINBAD theory of how the cerebral cortex operates is correct, or some other theory that shares its externalism, isomorphism, and psychofunctionalist aspects. But if empirical fortune goes SINBAD's way (so far, so good), teleosemantics need no longer be hostage to Swampman intutions. More ambitiously, this mere demonstration in *principle* that Swampman need not trouble teleosemantics should persuade some naturalistic philosophers to give Millikan, Dretske, et al. another look.

References

DESTEXHE, A. (2000). Modelling corticothalamic feedback and the gating of the thalamus by the cerebral cortex. *Journal of Physiology (Paris)*, *94*, 391-410.

DRETSKE, F. (1996). Absent Qualia. *Mind and Language*, *11(1)*, 78-85.

JONES, J. P., & PALMER, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J Neurophysiol*, *58*, 1187-211.

KRIEGESKORTE, N., MUR, M., RUFF, D., A., KIANI, R., BODURKA, J., ESTEKY, H., TANAKA, K., & BANDETTINI, P., A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *60(6)*, 1126-41.

LOGOTHETIS, N. K., PAULS, J., & POGGIO, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*, 552-63.

MCCORMICK, D. A., & BAL, T. (1994urrenTO, B6 0 12 0 3.216064 Tm F1.0 inferio a2sanistsQ q 1 0 0 1 7