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 in the tank. Dretske expects us to share this intuition, and invites us to extend it to Swampman, given the plausibility of his indicator semantics for biological systems. My argument will take a similar form.

My strategy will be three-pronged. First, I will appeal to a thoroughgoing, independently motivated externalism. Having never encountered her, Swampman lacks a concept of Davidson's mother. That much is generally accepted. Therefore if all mental representations, in order to play

[1] The paper is a bit over 3000 words as written here, but can be presented in 25 minutes due to the use of slides.

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 formation with the same spatial structure does not because it lacks the relevant teleology.

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.

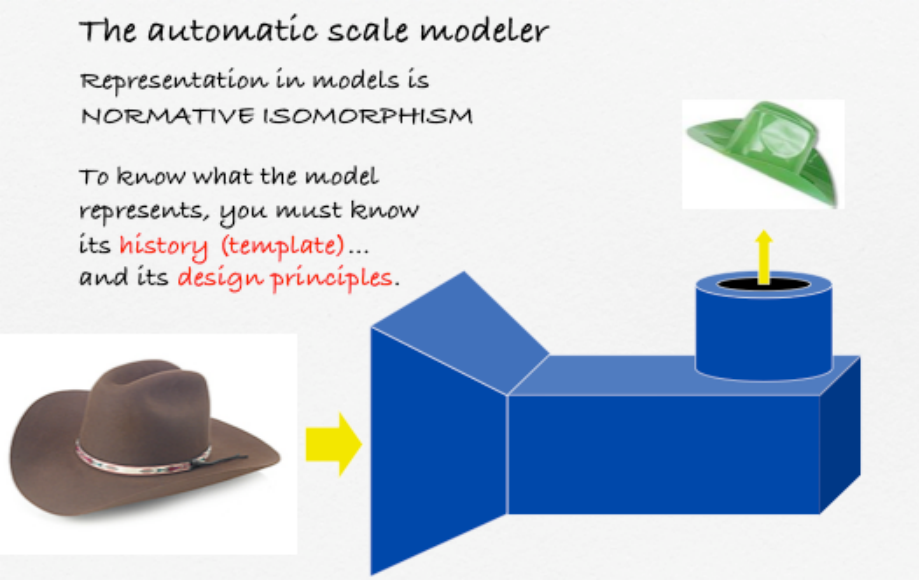

Figure 1: The automatic scale modeler

[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.

The SINBAD theory says that the cerebral cortex models regularities organized around sources of mutual information. Because of the details of the cortical learning mechanism, pyramidal cells in the cortex will naturally tune to SOMIs. Due to the same mechanisms, different cells will tend to choose different SOMIs - and these cells develop links that mirror the regularities that exist among those SOMIs. (These connections can mirror nonlinear relations involving multiple variables, which makes the SINBAD machine particularly powerful in modelling environmental regularities.) The connections among cells that develop in this way are analogous to the gears in the solar system model; they're what makes covariation in the network mirror covariation in the

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 (Logothetis et al., 1995; Kriegeskorte et al., 2008). Combined with the analysis of model representation as normative isomorphism, plus a demonstration that SINBAD isomorphisms are indeed normative (the model-building machine is designed by evolution), this all adds up to global externalism. What does this particular cell or cell group represent? The answer is determined by that cell or cell group's *template*.[4] Since Swampman's cells lack templates, they have no determinate denotation, just like a random pattern in a piece of toast doesn't denote Mary, mother of Jesus (photos also require templates!), and a blob resembling Elvis spat out by a malfunctioning automatic scale modeler doesn't represent Elvis. This applies equally

[3] Tuning to SOMIs carries with it a number of important benefits compared to other neural-network-type mechanisms, especially with respect to ease of learning.

[4] For details on how this determination goes, see [reference omitted for blind review.]

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 different optics and/or different photoreceptive equipment, a SINBAD network would develop a very different 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 "tweaks" to the fundamental SINBAD mechanism to encourage a particular trajectory to, e.g., visual or linguistic development.

[7] Actually, there is another necessary condition on model representation, which is that the model be subject to appropriate uses – see section 5.

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 distinguished from the truth-evaluable content that we normally ascribe to mental states. If we were to scan Davidson's brain while he were engaging in normal behaviour, we would be able to pick out the aspects of his cortical SINBAD model that he was using at any one moment. Further, by tracing how those aspects of the model influenced his behaviour, and how they were influenced by environmental inputs, we could figure out which objects in the environment Davidson was applying his SINBAD structures to. These would generally match the structures that they were supposed to apply to, although (of course) there would be exceptions. Davidson would sometimes mistakenly apply his model of a beech tree to an elm. In that case, Davidson would harbour false thoughts.

While Swampman couldn't harbour false thoughts in this way, we could still make the same determinations concerning how he was applying his internal structures to the world. Further, given a particular environment, we could figure out how he was *disposed* to apply his internal structures to the world. Where those structures were *accurate*, we could pragmatically assign contents to them. If it's most accurate about Davidson's house, and he's disposed to apply it to Davidson's house, then that can be its pragmatic content. If it's accurate about water, then that can be its content, etc.

Where one of Swampman's internal structures is equally accurate (and inaccurate) about multiple objects, then we have no reason to assign one as its content rather than another. By contrast, with Davidson we can refer back to structure's template to discover that the SINBAD cell in question represents elms. Therefore he has these particular false beliefs about elms, rather than these (complementary) false beliefs about beeches. There are no grounds for such decisions in the case

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 this world rather than a wildly inaccurate model of very bizarre world, other than the pragmatic one that he happens to be here with us. This should make it abundantly clear that Swampman's mental states have no proper truth evaluable, truth conditional content.

Returning to earth, we can explain Davidson's successful behaviour in two ways: 1) by the fact that his SINBAD models are reasonably isomorphic to a large swath of the world and its contents, and 2) by the fact that the world has imprinted itself on his internal model-making machine (according to the SINBAD principles) in order to create that isomorph. While we can't explain Swampman's behaviour in the second way[9], the first is still available to us. And the first can be tied to a pragmatic notion of content, allowing us, superficially at least, to explain Swampman's behaviour in much the same way that we explain Davidson's. He opened the door because he expected to see his mother in there; and he wanted to see his mother. Contents by fiat for Swampman; real contents for Davidson.

5. SINBAD in use, and the third prong

Before I present the third prong of the argument, which will show that Swampman can have perceptions, beliefs, and desires that lack truth evaluable contents, I need to fill in the rest of the SINBAD theory: how the SINBAD model gets *used*.

It's useful to have a concrete example here. Consider a model of dual-knob kitchen sinks. Such a sink contains a number of different variables, like the radial positions of the knobs, which knob is

---

[8] Compare Swampwoman, formed in a similar way to Swampman, but with a startling similarity to Ruth Millikan. However, Swamwoman's "brain" bears a much weaker resemblance, to Millikan's brain than Swampman's does to Davidson's. In fact, Swampwoman is behaviourally rather dysfunctional due to a poor isomorphism to Earthbound regularities. One is not inclined to attribute large numbers of false beliefs about particular natural kinds and individuals to Swampwoman; rather, her "concepts" are not determinately about anything. By contrast, if Ruth Millikan's brain were scrambled to match Swampwoman's, one would be much more inclined to attribute to her large numbers of false beliefs about her family, her cats, her house, water, and educational institutions.

[9] We therefore lose access to Dretske's dual explanandum strategy for the explanatory relevance of content – but that's another story.

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" for each of the sink variables, and enter into relations that mirror the covariation relations that obtain in the sink.

Because the relations of covariation in the sink are mirrored in the model of the sink, the model can be used to fill in missing information. Suppose that for some reason you cannot touch the tap water, and so are ignorant of the water temperature, but that you know the radial positions of the knobs. In order to remedy your ignorance, you could make the knob position stand-ins correspond to the positions of the knobs, and then just read off the water temperature from the stand-in for temperature at the tap. Since the model mirrors the covariation relations in the sink, imposing particular values on some of the model's variables will force the remaining variables to adopt values consistent with the covariational structure of the sink.

Filling-in in dynamic models also allows them to be used in guiding action. For example, suppose that I wish the water coming out of the tap to be at a particular temperature and to flow at a certain rate, but I do not know what positions to put the knobs in to obtain that temperature and flow. I may simply consult my model in order to find this out. I just make the stand-ins for temperature and flow correspond to the values that I want, and then I read off the knob positions from *their* stand-ins in the model. These stand-ins correspond to how the knobs *ought* to be, given my need for that temperature and flow. I can then use this information to turn the knobs to the appropriate positions, and get the temperature and flow that I want.

Now suppose an organism has an internal dynamic model of its environment, and that it "uses" its internal model in these two ways. In this case, the "uses" correspond to two different functional or causal roles that model can occupy. When the dynamic model is playing a causal role such that it is sensitive to its environment, and part of the model is made to correspond to the environment by receiving information from the organism's sensory receptors, the rest of the model will come to correspond to the state of the environment through the process of filling in missing information. In this functional mode, the model is supposed to correspond to the current state of the environment.

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), the state of the sink that is consistent with a high satisfaction. That is, instead of starting from a desired state of the tap variables, the internal model starts from a current need and a high level of satisfaction. It then asks itself, in what state should the sink be in order to be consistent with these values of satisfaction and need? It then "infers" the desired state of the tap variables by filling-in, and from there, infers the desired state of the knobs. (And then moves the knobs, of course.) In this functional mode, the model must be at least partially cut off from the environment, i.e. not be receiving complete information about the current state of the world. Rather, it is supposed to come to correspond not to how the world is, but how it *ought* to be, given the organism's needs.

In the brain, switching between these two functional modes is accomplished (in part) by mechanisms present in the thalamus (McCormick and Bal, 1994; Destexhe, 2000).[10] When a SINBAD cortical region is in "indication" mode, its activity is supposed to correspond to the way the world is. When a SINBAD cortical region is in "action-guiding" or "direction" mode, its activity is supposed to correspond to how the world ought to be, given the organism's needs. You will already have noticed that these directions of fit with the world match those that characterize beliefs and desires. Indication mode corresponds to the causal role characteristic of occurrent belief, more properly called judgement. Direction mode corresponds to occurrent desire. (One would imagine that there would be a third mode - confirmed to be present in an uncontrolled form

[10] These are the same mechanisms that are responsible for cutting off your awareness of external stimuli when you are asleep and dreaming. Their role also helps explain the fact that there are ten times the number of projections going back to the thalamus from the cortex than going the other way around, despite the traditional view of the thalamus as a mere waystation for information proceeding from the sensory periphery to the cortex.

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.

[12] Implicit beliefs are something different; they are contained in the relational structure of the network itself. For example, the kitchen sink network has the implicit belief that the knob variables are related by a certain function to the temperature at the tap.

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. (1994). Sensory gating mechanisms of the thalamus. *Current Opinion in Neurobiology*, *4*, 550-56.

PAPINEAU, D. (2001). The Status of Teleosemantics, or How to Stop Worrying about Swampman. *Australasian Journal of Philosophy*, *79(2)*, 279-89.