

DOI: 10.1017/psa.2025.10106

This is a manuscript accepted for publication in *Philosophy of Science*.

This version may be subject to change during the production process.

What really lives in the swamp? Thought experiments and the illustration of scientific reasoning

Andrew Richmond

Rotman Institute, Western University, London, Ontario, Canada.

arichmo8@uwo.ca

Funding information: None

Conflicts: No conflicts of interest to declare

Acknowledgements: Thanks to audiences at Columbia University, Western University, and the Canadian Philosophical Association for feedback, and thanks to André Curtis-Trudel, Linus Huang, John Morrison, Kate Pendoley, Jason Winning, and the EMRG Lab for helpful discussions. And thanks to a few anonymous reviewers for helpful comments.

Abstract

I use Swampman to illuminate the role of thought experiments in philosophy of science. Against Millikan and others, I argue that even outlandish thought experiments can shed light on science and scientific kinds, so long as we understand them as *illustrations of scientific reasoning*, not *examples of scientific kinds*. The logic of thought experiments, understood as illustrations, is analogous to the logic of common experimental paradigms in science, and allows Swampman to survive teleosemantic objections. So, in reviving Swampman, I also provide a framework for understanding how, why, and when thought experiments are informative about science and scientific kinds.

*If a random quantum fluctuation somehow created an iPhone
SE out of thin air it would still cost \$579. Checkmate Marxists*

— Jonathan Weisberg, on Twitter¹

1. Introduction

The philosophical bestiary is a surreal place. You can't see it all in one go, but on any given trip you're likely to visit, among other things: disembodied minds (Descartes, 1984) and envatted brains (Putnam, 1992); regretful vampires (Paul, 2014) and unconscious humanoids (Chalmers, 1996); perfectly choreographed simulations of brain activity performed by entire nations (Block, 1978) or by single, dedicated individuals (Searle, 1980); intergalactic duplicates of both earthly creatures and the psychology textbooks describing them (Egan, 2014); and, if you visit at the right time of day, 'Kimus' making their confused pilgrimage to the red of the sun, leaving their (just as adorably-named) predators behind (Pietroski, 1992).

It is, for philosophers of science, not at all clear what we're supposed to learn from the creatures that populate this menagerie. It's easy to think they are too unrealistic or fantastical to shed any light on the real world, or on real science. But they won't seem to go away. Consider how often *Swampman* sightings are still reported in respectable journals, and by intelligent people — most of whom (I assume) are not believers in garden-variety monsters, ghouls, or cryptids (Kim, 2021; Peters, 2014; Porter, 2020; Schulte, 2020; Sebastián, 2017; Tolly, 2021). Swampman is an especially odd case, for a couple reasons. First, many philosophers think Millikan killed Swampman dead pretty much the moment he rose from the swamp (Millikan, 1984). And second, her argument seemed to undermine not only Swampman, but *any* fantastical thought experiments like him, showing why they are irrelevant to our understanding of science.

This paper will develop an account of thought experiments to show how creatures like Swampman *can* be relevant to our understanding of science. In §2 I'll review Swampman, along with Millikan's argument against him, which has become the teleosemanticist's stock argument. In §3 I'll distinguish between the way Millikan and company conceive of Swampman — as an *example of a scientific kind* — and another way we might think of thought experiments — as *illustrations of scientific reasoning*. I'll show that illustrations don't rely on the assumptions Millikan criticizes, and I'll reinforce this by comparing the logic of illustrations with some experimental paradigms in cognitive science. In §4 I'll show that Swampman can be fruitfully

¹ (He's joking.)

understood as an illustration. So understood, he is untouched by standard objections and has interesting things to teach us about cognitive science. I'll conclude, in §5, with some implications for philosophy of cognitive science and philosophy of science more broadly.

2. Teleosemantics and the Swampman counterexample

Teleosemantics is easy enough to summarize at a high level: *representation* is a scientific kind, and what makes something a member of that kind is that it has a certain sort of selection history.² The view is applied to representation as it appears in folk psychology, philosophy, and cognitive science, but I'll focus on the cognitive scientific case, where authors like Neander (2017) and Shea (2018) use teleosemantics to understand the scientific role of the kind *representation*. It's worth mentioning that there are similar views of *computation* in philosophy of cognitive science (Milkowski, 2013; Piccinini, 2015), and of course selectional accounts of *function* that are important for cognitive science, even if they're mostly discussed in the context of biology (Egan, 2022; Garson, 2019; Neander, 1991; Wright, 1973). My discussion will bear on these views as well, but I'll focus on teleosemantics.³

Swampman is also pretty easy to describe: imagine a creature physically identical to Donald Davidson, but with no selection history. E.g., imagine he's created by a freak chemical reaction when lightning strikes a swamp. Of course (the argument goes) this creature would have representations. He would see, and plan, and wonder how he ended up in the swamp (being identical to Davidson at the moment of his creation, he will come into existence with the thought that just a moment ago he was sitting in a seminar room). We can ask him to deliberate over a career change: keep prowling swamps for a stable wage and good benefits or try for a riskier but potentially more fulfilling career as a philosopher. He can weigh the pros and cons, the state of the job market, and the possibility of finding himself redundant if swamp-prowling is automated. And, since he is a perfect copy of Davidson, he will give thoughtful answers to all these questions. The Swampman argument points to this paradigmatically representational activity and concludes that Swampman is a counterexample to teleosemantics: he has representations, but no selection history.⁴

² I'm ignoring teleosemanticists who appeal to non-selectional functions, like forward-looking functions (Nanay, 2014). They are not Swampman's target, or mine.

³ I'm setting aside teleosemanticists concerned solely with *non-scientific* notions of representation. My purpose, like Shea's and Neander's, is to illuminate representation as it is understood in cognitive science — not to compare the scientific notion of representation to 'genuine' representation of a deeper kind.

⁴ For early versions of Swampman, see Davidson (1987), Millikan (1984, p. 93), and Boorse (1976).

This is a tidy argument, but it was met pretty much immediately with what many philosophers see as a conclusive response. Scientific kinds are more sophisticated than the Swampman argument supposes. They are more like what Millikan calls *real kinds*:

Real kinds I define as groups over which a variety of relatively reliable inductions can successfully be run *not accidentally but for good reason*. The essence of a real kind is whatever accounts for its instances being alike. (Millikan, 1996)

Why should we think of scientific kinds this way? Because this is what science needs; kinds held together by mere similarity don't serve its purposes (Garson, 2022, p. 464; Neander, 1996, p. 120; and Shea, 2018, pp. 22, 28–29, are especially clear about this motivation). And creatures as fantastical as Swampman don't count as members of the same *real kinds* as actual organisms: Swampman is similar to representational systems, just like he's similar to members of the kind *homo sapiens* — but not for good reasons (Millikan, 2010; Neander, 1996). The reason that real representational systems (like you, me, and my pet bird) are similar is that we have similar evolutionary histories. The reason Swampman is similar to us is that philosophers wanted to describe something that looked as much like us as possible. So Swampman has some of our features, but not for the same reasons we have them, and therefore he does not belong to the same kinds as us — at least not *real kinds*, which are held together by something deeper than mere similarity. *Representation* is such a kind, so Swampman's states are not examples of the kind *representation* any more than he is an example of the kind *homo sapiens*. And so Swampman's lack of selection history can't tell us whether representation requires selection history.

It's important to note that this argument doesn't hinge on Millikan's account of kinds as induction-generators. Induction is not the only scientific task, nor the only task that kinds can serve. It is uncontroversial that science involves many and sundry tasks like modeling, communication, and understanding, among others (cf. Waters, 2019). The deeper point, elaborated especially by Shea (2018, pp. 28–29), is that our intuitions about which kinds Swampman belongs to should be overridden by facts about what the kinds must be, to play their scientific roles. To derive anything about a kind from an example of it, we need reasons to think our supposed example really is an example of the kind, and those reasons must be *grounded in the scientific role of that kind*.

What would such a reason look like? One candidate is that the best explanations of Swampman would attribute representations to him. Surely that makes it plausible that he really does represent (Neander, 1996, p. 123). But the same sort of response can be given here. Like other scientific kinds, *representation*'s role is to help capture and explain worldly patterns. And Swampman isn't an instance of the worldly pattern cognitive science tries to capture and explain when it uses the kind *representation*, any more than he's an instance of the pattern biology tries to capture and explain when it uses the kind *homo sapiens*. Again, scientific kinds must be defined by the features they need to serve their scientific roles — that does not include the features they *would* need to serve an *imaginary* science (let alone a fantastical science tasked to explain Swampman-style abiogenesis). The fact that an imaginary science in a fictitious world would explain Swampman in human and representational terms doesn't establish him as a member of the worldly patterns that kinds like *representation* and *homo sapiens* are used to capture and explain. And if we haven't established that, we haven't established that Swampman has representations, so his lack of selection history cannot tell us whether representation requires selection history.⁵

To sum up, the Swampman counterexample tries to derive something about the nature of representation from an example of that kind. But if we want to derive the nature of a scientific kind from examples of it, we first need to argue that our example really *is* an example of the relevant kind. And teleosemanticists, beginning with Millikan, have made a good case that, because of his fantastical nature, we are unlikely to see a convincing argument that Swampman really is a representational system — at least if we're committed to taking *representation* seriously as a scientific kind, and understanding it in terms of its role in real science (cf Millikan, 2010, p. 79; Papineau, 2001).⁶

I think the teleosemanticists are right: trying to derive the nature of representation from Swampman, understood as an example of a representational system, is a mistake. And I think

⁵ There is also a more straightforward argument: we often explain things in terms of kinds they don't belong to. We model traffic as a *fluid*. We explain disinformation by describing it as a *virus*. So assume we *would* explain Swampman by attributing representations to him. That leaves it open whether the attributions would be literally true, i.e., whether he literally has representations. But that's what the argument at issue was supposed to establish.

⁶ You could characterize the arguments in this section as instances of Häggqvist's (2009) *biting the bullet* and *irrelevance defenses* against thought experiments, with thought experiments understood on the logic of *alethic refuters* or counterexamples (Cohnitz & Häggqvist, 2018; Sorensen, 1992). The point of this article is to describe another logic for thought experiments that — when sound — undermines those defenses by illuminating exactly how thought experiments can be relevant to our understanding of science and scientific kinds without functioning as counterexamples.

they're right about why: that would require an argument that Swampman really is an example of a representational system, and such an argument — especially one based on the scientific role of the kind *representation* — doesn't seem to be forthcoming. But I want to reconsider the kind of argument Swampman is supposed to be. I'll argue that Swampman should not be understood as an attempt to derive the nature of a scientific kind from an example of it, but as an attempt to do exactly what the teleosemanticist urges: to probe scientific explanations, the role that the kind *representation* plays in them, and the features the kind must have to play that role. We can use Swampman to probe scientific explanations just like we use cardboard cutouts to probe frog prey detection, without any commitment to those cutouts being real examples of the kind *frog prey*. The next section will spell this out in detail.

3. Examples and illustrations

In §3.1 I'll develop a simple thought experiment, analogous to Swampman, that doesn't make problematic assumptions about kind membership. I'll call this sort of thought experiment an *illustration*.⁷ In §3.2 I'll compare the logic of illustration with the logic of some experimental paradigms in cognitive science. Partly, this is building to the argument in §4 that Swampman is best understood as an illustration. But it also stands on its own as an account of the way that thought experiments — even fantastical ones — can serve philosophy of science.

3.1 An example of illustration

Imagine a student in a physics class who expressed the following misunderstanding:

“Explanations in quantum mechanics appeal to *observers*, so quantum mechanical effects rely on conscious observation. Therefore [some woo].” Though this is wrong, *observation* is a difficult and contested notion in physics, and it might be impossible to point the student to a passage in the textbook with a clear definition of the term. A better tactic would be to show the student that quantum mechanical (QM) explanations work, predict, and explain experimental results even if any potential conscious observers are, say, looking away from the experiments. If the student's misunderstanding persisted (“But we set up the device that does the observation, and *we're* conscious”), you might respond with a more extreme hypothetical: a universe without conscious observers at all. We can imagine a physics lab popping into existence in that universe, arranged so that it sets into motion a classic QM experiment. Our QM explanations, models, and

⁷ Not to be confused with the “illustrative thought experiments” of Cohnitz & Häggqvist (2018), which are intended to illustrate a *theory* by describing its implications. I aim to illustrate a *target system* by describing how it functions.

predictions, including everything they say about observation, would still succeed there, and in all the same ways they succeed in the actual world.

All we're doing here is paring away an irrelevant feature (consciousness) to show that it is, in fact, irrelevant to our explanations. We're hoping the student follows us in an inference that goes something like this: if QM explanations go through just as well, and in all the same ways, whether consciousness does or doesn't exist, then those explanations must not rely, for their success, on consciousness. And if *observation* is defined by the features it must have to serve those explanations, consciousness won't be part of its definition.

Now, does it matter that consciousness actually exists? That the universe we've described, with a physics lab but no consciousness, is imaginary and fantastical? Such a universe is surely not an instance of the same worldly pattern as our own, which developed according to dynamical laws — not by the random coming-into-existence of complex and apparently goal-directed things like physics laboratories. A world with physics labs but no consciousness might even be *impossible*, on certain (niche) views of consciousness (Goff et al., 2022; Tononi et al., 2016). But none of this matters because you're not using the thought experiment, in the first place, as an example of the kind *observation*, and then deriving the nature of the kind from that example. In the terminology I'll use, you're using the thought experiment to *illustrate* a type of scientific explanation, by paring away a feature to show that it is irrelevant to the success of that type of explanation.

To return to the main lesson of Millikan's response to Swampman, note that in the QM case we're starting with actual scientific explanations, and we're asking how they work in their actual contexts. It's just that we answer this question by paring away features of those contexts and checking the consequences for the explanations' success. That does involve a hypothetical non-conscious world, but we're not starting from a claim that the non-conscious world is a member of some kind, and nor are we saying that, because we would explain it in terms of some kind, it must be a member of that kind. Instead, we're showing that our actual explanations would be *just as successful, in all the same ways*, if consciousness was absent. And we're drawing the natural conclusion: the success of those explanations does not depend on the presence of consciousness. Kinds don't even come into the picture until a further step: if we agree with the teleosemanticist that the nature of a scientific kind should be determined by its scientific role, the kind

observation should not be defined in terms of consciousness because consciousness is irrelevant to the kind's role and the explanations it serves.

3.2 *The logic of illustration*

The logic of illustration is fundamentally no different than the logic you might use to understand how any system performs any task, from how primates develop emotional capacities (Harlow, 1959) to how cooks make a good *cacio e pepe* (Bartolucci et al., 2025): pare away parts and ask whether the system is still able to perform the task to the same degree and in the same way. If it is, then the part you removed likely wasn't contributing to the task in the first place. For example, if you want to understand how a can-opener works, you can determine that its color properties are irrelevant by paring them away (either really or imaginatively) and seeing whether you have impeded its ability to open cans. The same goes when the system is *science* rather than a kitchen utensil, and the task is *explanation* rather than opening cans.

But I've aligned myself with the side of the debate that thinks we should take scientific explanation more seriously, and I don't think anyone will be satisfied by a comparison between explanations and can-openers. So I want to flesh out the logic of illustration by comparing it to two experimental paradigms in cognitive science. These comparisons will reinforce the point that illustrations don't make problematic assumptions about kind-membership. And they will support the *legitimacy* of illustration (at least, insofar as the paradigms I compare it to are legitimate) as well as introducing some challenges that illustrations must face.

To set the stage for these comparisons, we can break the logic of illustration down into three components.

1. There is a *broad explanatory target*: we're trying to understand how a form of explanation (e.g., QM explanation) successfully explains events and patterns.
2. We *narrow this target* to ask about a feature of the explanations (e.g., the consciousness of observing devices) — does that resource contribute to the explanations' success?
3. And we *probe* this question by paring that resource away (e.g., imagining an unconscious lab) — what would happen to the success of the explanations if that resource was missing?

How does this compare to typical cases of cognitive scientific reasoning? One characterization of cognitive science is “the study of how agents perform tasks” (Mekik & Galang, 2022, p. 2), and specifically how they use different resources to perform them — both internal resources (like

neural structures or activity patterns) and external ones (like features of the environment).⁸ To answer these questions, cognitive science routinely removes resources and examines the effect on task performance — just what I’m claiming illustrations do. In the rest of this section I’ll compare illustration to one cognitive scientific paradigm that removes internal resources, and one that removes external resources.

Let’s start with the former. Ablation studies investigate the role of a brain area in some task by either ablating that area, or finding organisms in whom it has been ablated naturally, e.g., by strokes or railroad spikes (Damásio et al., 1994; Salvalaggio et al., 2020). Classical work found, among other things, that “bilateral lesions to lateral occipital–temporal cortex could lead to impairments in recognizing objects but no difficulty performing grasping and reaching movements to the same objects,” prompting the inference that the lesioned area was used in the former task, but not the latter (Mahon & Hickok, 2016, p. 942). Current work creates more carefully targeted ablations (Liu et al., 2019; Zhang et al., 2021) or temporarily disrupts activity in a brain area, assuming that the area is incapacitated by the disruption (Weissman-Fogel & Granovsky, 2019). But the inferences have the same form as always: if an area or its activity can be eliminated without affecting task performance, it must not have been used to perform the task in the first place (von Eckardt Klein, 1977; Bickle et al., 2019, p. 34).⁹ That logic comes with important caveats, which I’ll discuss shortly, but it bears comparison to the logic of illustration:

1. There is a *broad explanatory target*: we’re trying to understand how an organism (e.g., a human being) successfully performs certain tasks.
2. We *narrow this target* to ask about a feature of the organism (e.g., a particular brain area) — does that feature contribute to the organism’s performance?
3. And we *probe* this question by paring that feature away (literally ablating it) — what happens to the organism’s performance when that feature is missing?

There were two things I wanted to do with these comparisons. First, I wanted to make it clear that this logic does not require any illicit assumptions about the kinds our target organisms belong to. To see this, note what the experiments conclude: just as an illustration probes the resources *real* scientific explanations use (not just the resources the imaginary one uses), ablation

⁸ *Use*, here, is not an intentional notion. Agents can use resources intentionally, but when they’re using a part of their hippocampus to navigate, we’re dealing with a purely functional notion (cf. Baker et al., n.d.).

⁹ Many ablation studies *do* find an effect on task performance and conclude that the ablated area did contribute to the task. But the illustrations I discuss are ones where removing a feature has no effect on explanations, so I’ll focus on the corresponding subset of ablation studies.

studies conclude that *intact* organisms of a certain species do or don't use a brain area to perform a task (not just that the *ablated* organism does). So the conclusions aren't based on an assumption that the ablated organisms are members of the kind we're trying to draw conclusions about — ablated organisms are, by definition, not intact organisms. In fact, they often aren't even the same species as the organisms we're ultimately trying to understand (e.g., Weiskrantz et al., 1974, p. 709), and there are even attempts to draw conclusions about human brains from ablations in artificial neural networks (Lillian et al., 2018). Clearly, the conclusions about intact organisms of a particular species do not derive from any assumption that the ablated system is *also* an intact organism of that species.¹⁰ They derive, instead, from an assumption that the two organisms are performing the task the same way, i.e., using the same brain areas or structures. So it would miss the point entirely to apply the *real kinds* response here and say, "The ablated organisms do not belong to the kind *intact organisms*, so we cannot draw any conclusions about intact organisms from them."

The second thing I wanted to do was highlight challenges that apply to ablation experiments and maybe, by extension, illustrations. The main challenge to ablation experiments concerns the possibility that an organism learned to perform its task in a new way, or that its brain somehow compensated for the missing area. That is, our assumption that the organisms perform their tasks the same way might be mistaken. Maybe the ablated area *is* used in the task, but it's redundant. Or maybe another brain area took over its role in the ablated organism. The point is that *similar task performance* does not strictly entail *similar resources used* (e.g., Barsalou, 2016, p. 1128). I won't rehearse the many examples of this here. Suffice it to say, ablation studies have a difficult problem: they have to make the case that the organism under study really is doing the task the same way as the target (intact) organisms. In practice, this means ruling out plausible learning or compensatory mechanisms, e.g., by eliminating the necessary time for learning or plasticity, or showing that brain areas that might be expected to compensate don't actually change their activity patterns after the ablation. To make a plausible argument from an ablation study you must do this convincingly, and the same goes for illustrations. If you use the illustration I described in the QM case, you need to rule out plausible ways the explanations might have

¹⁰ Of course, the ablated organism has *some* kinds in common with the intact one, just as Swampman does with human beings. The point is that the reasoning doesn't rely on the two organisms sharing *the kind we're drawing conclusions about*: intact members of a certain species.

succeeded even if they originally *did* rely on the consciousness of the observer. Perhaps some concept does extra explanatory work, or concepts change roles and contribute differently.

The other type of experiment I promised to discuss is one that pares away an organism's *external* resources, namely, features of its environment. There is a particularly clear description of this experimental paradigm in a paper that teleosemanticists are familiar with, investigating prey capture in frogs (Lettvin et al., 1959). The set-up of that paper tells us that a frog "will starve to death surrounded by food if [the food] is not moving" (p. 1940). Eliminate motion, and prey capture is affected. So motion seems to be one of the environmental features prey capture depends on. But the frog "will leap to capture any object the size of an insect or worm, providing it moves like one. He can be fooled easily not only by a bit of dangled meat but by any moving small object" (p. 1940). Eliminate any of the prey's features aside from its rough physical dimensions and movement patterns, and prey capture isn't affected. So prey capture must not depend on those features. The logic is even more transparent here than it was with ablation:

1. There is a *broad explanatory target*: we're trying to understand how an organism (e.g., the frog) performs certain tasks.
2. We *narrow this target* to ask whether the organism relies on a particular resource (e.g., a feature of the environment) — does that resource contribute to the organism's performance?
3. And we *probe* this question by paring that resource away (either stripping a feature from the environment or constructing an environment lacking the feature) — what happens to the organism's performance when that resource is missing?

Lettvin et. al. used stimuli that had various features in common with flies, a typical prey item for the frog. And they recorded from a number of nerve fibers assumed to drive prey capture behavior, to see which stimuli caused a change in their response and which didn't. Eliminating all the physical features of a 'fly' except its rough size and shape did not stop the fibers from responding, but eliminating its movement relative to a background did (p. 1945). The point is that, assuming that these nerve fibers' response is what drives prey capture, the physical features of a fly aside from its rough size, shape, and movement patterns do not contribute to the prey capture task.

To be fair, these experiments don't exactly represent the state of the art. Neander (2017) brings out the increasing complexity of anuran prey capture research. But much of that research

follows the same logic, with the essential finding being that “the relevant visual discrimination in an unconditioned toad is largely unaffected by features not captured by [a small number of] dummy stimuli” — the famous cardboard cutouts of worms and flies (Neander, 2017, p. 104). That finding is so important because, by telling us which features can be eliminated with no effect on prey capture, it tells us which features are and aren’t used in that task.

How would the *real kinds* response fare against this body of work? About as well as it did against ablation studies. Here we’re using non-naturalistic environments, including some wildly unrealistic ones — cardboard cutouts standing in for worms and flies — to probe the way that anuran prey capture works in real, natural, environments. We might object that prey capture works differently in the two environments, but these objections would appeal to facts about how prey capture works in those environments, not kind-membership. They would not say, “Well, these are very nice experiments, but unfortunately, you’ve made a terrible error: the non-naturalistic environment isn’t of a kind with the naturalistic one! Shame you spent so much time on this study; it’s worthless.” The reason that sounds so ridiculous is that the work in question puts no evidential or epistemic weight on the experimental stimuli (cardboard cutouts) being examples of the kind of stimuli we’re trying to draw conclusions about (real prey), but only on the fact that the frog is using the same resources in the non-naturalistic environment as in the naturalistic one. I’m not trying to mock Millikan here; she didn’t give the *real kinds* response to this sort of experiment. The point is that she was right not to: it would have been a mistake to think that the logic of these studies — and, by extension, the logic of illustration — relied on the experimental situation (non-naturalistic prey capture) being an example of the kind (naturalistic prey capture) that we’re ultimately drawing conclusions about.

This comparison raises complications similar to the ones we saw with ablation studies: it’s possible to argue that prey capture works differently in naturalistic and non-naturalistic environments. Ecological psychologists have been making points like this for decades (Baggs & Sanches de Oliveira, 2024; Gibson, 1972; but see Shepard, 1984): in impoverished (experimental) environments, organisms may perform tasks differently than they do normally, when they have more environmental information to use. This is a harder case to make when we’re only eliminating one or a few features of the environment, but in any case, the task for the frog researcher is the same: to show that the *potential* differences in how prey capture works are not *actual*.

For a different example, consider Hartle & Wilcox's (2016) study of stereopsis. They were investigating binocular cues for depth perception, but, chasing down some unexpected patterns in the data, they realized that their participants had discovered artefactual *monocular* cues in the experimental stimuli. So Hartle & Wilcox found that they could eliminate binocular cues without affecting depth perception, only because they hadn't just removed binocular cues: they had added monocular cues for their subjects to use instead. That undermines any conclusion you might draw to the effect that binocular cues are not used in depth perception. This is just to say that if you pare away features of the environment to check the effects on task performance, and if you want to draw conclusions about how the task is performed in normal, non-experimental settings, you need a plausible argument that the 'paring' didn't induce the organism to perform its task in a new way, e.g., by accidentally introducing new resources, or causing the organism to use different strategies or processing than it does in naturalistic environments. Likewise, illustrations will need to show that *their* manipulations don't have unintended effects on the way scientific explanations work or introduce new resources for the explanations to use.

4. Swampman redux

I've described a way thought experiments can work without the kind-membership assumptions that teleosemanticists criticize. In §4.1 I'll show how the Swampman argument works on this understanding of thought experiments. As an *illustration*, Swampman escapes the *real kinds* objection for the same reason the experimental paradigms above do. In §4.2 I'll show what some other objections to Swampman look like when he's understood as an illustration.

4.1 Swampman as an illustration

This argument will sound familiar. It's supposed to. I'm trying to capture the argument that philosophers ought to be making when they talk about Swampman, which is different in small but essential ways from the argument that they tend to make, and that teleosemanticists then respond to. (Hacohen, 2022's rendition of the argument is similar to the one that will follow, but does not draw out the structure of the argument as I do here.) Let Swampman be generated in a swamp just as normal. We can imagine him wandering into a university building and seeing posters advertising Calls for Participants. We can imagine that he signs up for a study and lands in a psychology or neuroscience lab. Since he is a molecule-for-molecule copy of Donald Davidson, he will display the same cognitive capacities as Davidson. So the first major step of

the Swampman illustration is a simple disjunction: Swampman's capacities are either explicable or inexplicable.

I need to rule out some easy responses on behalf of the teleosemanticist here. Swampman will display the same capacities as Davidson only insofar as those capacities are described without reference to history, or to properties that rely on being the 'real' Davidson. But this is not a significant limitation. Cognitive scientists might study the accuracy of Swampman and Davidson's memory, but I doubt anyone would argue that what they mean by *accurate* is such that Davidson has the capacity to accurately recognize his mother, but Swampman doesn't because she didn't actually give birth to him. Those kinds of historically-inflected capacities might be interesting, but they are outside the remit of cognitive science and of typical representational explanations — my target here. So I take it as given that, as far as representational explanation is concerned, Davidson's and Swampman's cognitive abilities are the same, and the interesting role for selection history is as a *resource* to explain those abilities.

More specifically, when I consider the explicability of Swampman's capacities I will be considering capacities that cognitive science paradigmatically explains in *representational terms*, in ways that teleosemanticists take to support the selectional notion of representation. This includes, for example, explanations of navigation (Shea, 2018, Chapter 5), prey capture (Neander, 2017, Chapter 5), and the communication of resource locations (Millikan, 1984, Chapter 5), in addition to memory. It's those explanations, and their reliance on selection history, that stand to confirm or disconfirm teleosemantics. This will limit the generality of my conclusions. I will (thankfully) not be able to conclude that selection history is entirely irrelevant to cognitive science — just to representational explanation.¹¹

This also means that it's not Swampman's history that's up for explanation; we're not asking whether the way Swampman *came by* his capacities is explicable. We're asking whether

¹¹ Teleosemanticists might say that cognitive science has nothing to explain about Swampman: cognitive science aims to explain *successful* behavior, and without a selection history Swampman can have no ends to be successful with respect to (cf. Shea, 2018, p. 22). This would make things easy for the teleosemanticist, but it isn't a plausible characterization of cognitive science. Cognitive scientists aim to explain, e.g., the way you forage in your environment, or the fact that you make it from one place to another more often than chance would have it if you're given certain cues. These explananda do not disappear if we stop characterizing them as successes-in-a-selectional-sense. We can characterize them as simply patterns of behavior, or as successes in a non-selectional sense (compare fn. 2, and see Baker et al., n.d.; Mekik & Galang, 2022, for selection-agnostic notions of tasks and task success). So while selection history may be an explanatory resource, it is not a pre-condition for thinking scientifically about cognition. Louise Antony makes a similar point about biology, insofar as it's motivated by medical purposes: an oncologist would be just as interested in treating a patient whether they came from the Swamp or from Dallas (1996, p. 72). The question is what resources she would rely on, and whether they would include selection history.

cognitive science can or can't show how Swampman's physical organization supports capacities like navigation, memory, prey capture, etc. — whether it can or can't *reverse-engineer* those capacities. What would it be to accept the “can't” side of those disjunctions? If we sat Swampman down in the laboratory, would we be stymied-in-principle by his behavior? Would it be impossible to model the structure of his brain at the levels of grain that allow us to predict and explain his actions? Or would this project at least be less successful than it is with Davidson? There are no tricks up Swampman's sleeve — he's just another physical system. I don't see any way of denying that his various capacities would be as explicable as the capacities of any system. For anything he can do, there must be an explanation of how he does it.

The next step is to ask *how* we would explain Swampman's capacities. Would we have to use explanatory methods, strategies, or resources different than the ones we use to explain Davidson? I don't see any way of supporting that view either. We would observe Swampman the same way we observe organisms whose evolutionary history is unknown to us. We would see patterns in his behavior: a tendency to forage in his environment in a certain way; an ability to learn patterns in sets of stimuli; and so on. And we would investigate those behavioral patterns with questionnaires and response-time measurements, black-box models and eye-tracking experiments, computer simulations and fMRI data, circuit diagrams and electrode recordings, and so on. That is: we would apply the same explanatory resources that we do in cognitive science more generally. And we would take the same approach to building models of Swampman — including the use of representational notions.

To demonstrate this, imagine a meta-experiment: a single-blind trial where the participants are *two cognitive neuroscience labs*. We send Davidson off to one lab, and Swampman to the other. But we've given Swampman a shower and some clothes that aren't covered in swamp goo, so the scientists can't tell who is Swampman and who is Davidson. If we think that the success of cognitive scientific explanations relies on their targets' selection histories, we must think they would fail in some respect when applied to Swampman but would be successful in that respect when applied to Davidson. What respect could this be? The models we would construct would be equally predictive of each's behavior, as revealing of its neural basis, as useful in medical interventions. Nothing about Swampman, his swamp-brain, or his swamp-engagement-with-his-environment would seem to impede those projects, any more than our typical ignorance of an organism's evolutionary history impedes those projects (cf. Hacoen, 2022).

Again, keep in mind that we aren't talking about *all* projects cognitive science might have. A cognitive scientist might explain how an organism evolved a certain brain organization, and selection history is clearly relevant to that question. This is just to reiterate that what's at issue here are representational explanations, and the capacities they typically explain. There must be some goal the Davidson lab would achieve, and the Swampman lab wouldn't (or would only to a lesser degree) when it tried to explain, in representational terms, how Swampman's physical organization supports tasks like navigation or memory or perceptual discrimination. To come to the point of the Swampman illustration, I've suggested that there is no reason to think that scientists in the Swampman lab would fail in any way that the Davidson lab wouldn't. If that's the case, representational explanations must not rely, for their success, on selection history, because with selection history 'ablated' they would work just as well, and in all the same ways, as they do normally. So if we agree with the teleosemanticist that our account of a scientific kind should be determined by its role in scientific explanation, we should not define representation in terms of selection history for the same reason we should not define *observation* in terms of consciousness. Since there are no problematic assumptions here about kind-membership (the logic of the argument is the logic described in §3.2), the teleosemanticist's stock objection, the *real kinds* response, doesn't apply.

I've gone quickly, and it's possible that I've missed some explanatory goals that wouldn't be met when cognitive scientists explained Swampman (I'll return to this shortly). But my goal has just been to cast Swampman in a more compelling role than he normally plays, and to show that the usual teleosemantic response doesn't apply to him in this role. If representational explanations have goals that I haven't discussed, and that expose a difference in explanatory success between Swampman and Davidson, then the teleosemanticist would need to describe those goals, argue that they *are* goals of representational explanation as it is actually practiced, and show how representational explanations of Swampman's capacities would fall short of them. For now, I'm satisfied if I've got Swampman back on his feet.

Before I move on, what about the challenges we saw above for experimental paradigms with the same logic as illustration? First, we have to rule out *compensation*, where some other feature of Swampman or the lab explaining him does extra work when his selection history is absent. It's not clear what this would be. Does the lab that receives Swampman have to put extra emphasis on behavioral as opposed to brain data? Will that lab make additional modeling assumptions?

There doesn't seem to be a plausible compensatory mechanism, especially since, in the hypothetical experiment, the two labs don't know who got Swampman and who got Davidson. Second, we have to rule out any unintended effects of removing Swampman's selection history, and especially any resources that might *introduce*. But there don't seem to be any plausible worries here. Swampman will have had a very different day than Davidson, who woke up that morning in a bed, not a swamp. But that sort of difference doesn't seem to provide resources that could fill explanatory gaps left by Swampman's missing selection history or allow the Swampman lab to reach the same conclusions about their subject as the Davidson lab does by different means. (But see below for comments on Swampman's *short-term* selection history).

4.2 Other objections to Swampman

I want to show that understanding Swampman as an illustration, rather than an example, helps deal with some teleosemantic objections aside from the *real kinds* response. First, teleosemanticists might accept that our explanations of Swampman work just as well as our explanations of Davidson, in all the same ways, but argue that some broader explanatory goal would be undermined if we defined representation in non-selectional terms. E.g., it is sometimes suggested — or, more accurately: stipulated — that cognitive science is in the business of generalizing over species-kinds (Millikan, 1996, p. 109) and maximizing the reliability (Millikan, 1996, p. 108) or breadth (Neander, 1996, p. 123) of its generalizations. This, supposedly, makes selectional kinds necessary. But these are dubious characterizations of cognitive science as a whole and representational explanation in particular, neither of which just aim to generalize but to model, problem-solve, explain, and so on, as I noted in §2. And anyways, it is not at all obvious that non-selectional kinds would fail to support the necessary generalizations and inductions — something that is acknowledged even by teleosemanticists (Shea, 2018, p. 22). So this appeal to broader goals is not a plausible objection unless teleosemanticists can (as I described at the end of §4.1) argue for some conception of those goals and show how representational explanations would fall short of them if some or all of its target systems lacked selection histories.

Another common objection to Swampman comes from teleosemanticists who define selection to include learning, differential survival, and other short-term processes (e.g., Garson & Papineau, 2019; Millikan, 1984; Neander, 2017; Shea, 2018). These teleosemanticists think even Swampman has a selection history (his morning will include at least some differential survival,

and probably a bit of learning), so the Swampman illustration *doesn't* show that representational explanations work just as well in the absence of selection history. If we understand Swampman as an example of a representational system, the teleosemanticist's assent means he's no longer a counterexample to teleosemantics; we have to ask whether Swampman has representations when short-term selection processes haven't had a chance to act, if such a situation is even possible.

But if we understand Swampman as an *illustration*, we have other options. First, we can change the thought experiment so Swampman just has *less* of a selection history. This is easier than going back into lateral occipital-temporal cortex to ablate parts you failed to get on the first try: just go up a couple pages and send Swampman to the lab sooner. Nothing in the thought experiment relied on significant stretches of time, so this shouldn't change the results. And second, note that teleosemanticists who are enthusiastic about short-term selection history tend to accept that evolutionary selection is *also* relevant (e.g., Shea, 2018). A teleosemanticist who appeals to *only* short-term selection processes is as rare a sighting as Swampman. The short-termist objection, then, would be that although we have ablated the bulk of Swampman's selection processes, the remaining scraps (short-term selection processes) can serve representational explanation just as well as before, without any visible differences. This puts them in the unenviable position of having to explain why so much of their view is unnecessary: what are long-term selection processes doing in their account of representational explanation, if their absence makes no difference to representational explanation? This is especially damning when combined with the first response, so all that remain are arbitrarily brief selection processes.

Another common response points out that Swampman is (I don't know how to put this delicately) *just a bit ridiculous*. Leave aside the philosophical niceties, aren't the scientists laughing at us (e.g., Dennett, 1988, 2007; Millikan, 1996)? It's tempting to dismiss things like Swampman by just saying, "Bah! We're doing serious work here." But, as the study of frog vision makes clear, legitimate scientific inquiries also put their target systems in ridiculous scenarios made up of fake organisms. (Or consider the virtual realities that laboratory mice and flies spend so much of their lives in.) Just as in regular scientific experimentation, what matters is not the realism of the environment, but *that the deviations from reality are motivated*. As Douglas Mook puts it, one cannot just point out that an experimental setting is unrealistic and

reject it as externally invalid.¹² To judge an experiment there is simply “no alternative to thinking through, case by case, (a) what conclusion we want to draw and (b) whether the specifics of our sample or setting will prevent us from drawing it” (1983, p. 386). Cardboard cutouts are even less similar to flies than Swampman is to Davidson, but they are relevant to real prey detection because they are used in a chain of reasoning that is carefully designed to let us draw conclusions about real prey detection. For the same reasons, what matters is not Swampman’s realism, but that his specifics allow us to draw the conclusions we’re drawing via the chain of reasoning described in §3.

A more ecumenical response to the ‘Bah’ objection might bring Swampman down to earth, using real organisms to make a similar point. We might invoke organisms with evolutionarily novel traits that haven’t had a chance to be selected for yet (Peacocke, 2014; Peters, 2014; Porter, 2020; Walsh & Ariew, 1996). The problem with these cases is that they can be nit-picked to death. Sure, the trait is evolutionarily new, but is there some broader selected mechanism that confers content on it? Might it have derived content? Could it have an evolutionary pre-cursor of a *similar enough* kind for it to count as selected for? The advantage of Swampman is the same advantage cardboard cut-outs have over real, immobilized insects: he affords us control over exactly the features we want to manipulate, and keeps the investigation from being swamped by confounds.

A related worry is that the explanations of Swampman are only successful because we made him up to be as similar to Davidson as possible: they take advantage of *pre-existing* explanations of Davidson. That seems like sleight of hand. And it may be, if we’re arguing first of all that Swampman’s states are examples of the kind *representation*. Then his surface similarity to Davidson seems to trick us into thinking of him or explaining him in representational terms, regardless of his status as a representational system. But if we are, first of all, illustrating representational explanation rather than defining the kind *representation*, the logic is entirely different. Consider the frogs again. In order for us to make inferences about real prey detection, the stimuli *must* be similar enough, except in the features of interest, that the frog applies the same prey capture mechanisms as it does for real prey. The whole point is to “trick” the frog into triggering those same mechanisms, using clever stimuli that isolate particular features and

¹² For this reason, external validity is not (pace Sartori, 2023) an appropriate way to make sense of how at least some thought experiments, like this one, allows us to draw conclusions about the world.

therefore allow us to make inferences about those features' role in triggering the mechanisms. That's not sleight of hand; it's experimental design. So Swampman as an illustration is in no worse shape than a typical anuran vision experiment.

One last version of the 'Bah' objection might dismiss Swampman because he's a *thought* experiment, rather than a real experiment. Is it a problem that he's imaginary? It would be, if that meant he only probed our intuitions or invoked some kind of "voodoo epistemology" (Sorensen, 1992, p. 27). But the reasoning that an illustration asks you to undertake is more like prediction than intuition. *What would happen if Swampman walked into the lab* isn't mere intuition-mongering any more than *What would Mom do if she found out I got in a fight* or *How would this organism behave in its Normal environment* (Millikan, 1984)? As long as we can be reasonably confident in our predictions, these are unproblematic questions — even if Mom doesn't find out, and the situation remains imaginary.

5. Upshots

Let me issue a brief reminder, which will also give me a chance to recap the argument: I haven't been arguing that *since Swampman would be explained in representational terms, he represents*. As I discussed in §2, the teleosemanticist would respond (I think correctly) that this argument doesn't take the scientific role of the kind *representation* seriously enough. My argument has been that since representational explanations of Swampman would be just as successful as they are of cognitive science's paradigmatic targets, and in all the same ways, those explanations must not rely on selection history. That itself is revealing, but if we accept that a kind should be defined in terms of its scientific/explanatory role then we can take the further step to conclude that the kind *representation* shouldn't be defined in terms of selection history.

The significance of this re-framing is revealed when we return to the *real kinds* response and Swampman's distance from real representational systems. As I put Millikan's point above: (1) to derive the nature of a scientific kind from examples, we need real examples of it, or at least realistic ones, because (2) we need to probe the kind's role in real science. But (1) I'm not deriving the nature of a scientific kind from examples of it, but from a consideration of precisely (2) its role in real scientific explanations, and specifically the resources those explanations rely on. Moreover, I'm deriving its role in real scientific explanations using the same logic that scientists use to investigate the role of different resources in cognition, where they too make use of fantastical 'organisms.' This allows Swampman, as strange a beast as he is, to be informative

about scientific explanation — as long as we use him correctly, not as an example but as an illustration.

By now, I have taken up a considerable amount of your time trying to rescue a member of philosophy's ridiculous bestiary from extinction. Surely, I owe you some implications. I'll draw out two for philosophy of cognitive science, and two for philosophy of science more generally.¹³

Let's start with philosophy of science generally. The first implication is that, when we look at the sciences, we should embrace the task of *illustrating scientific reasoning*, rather than just, or first of all, *characterizing scientific kinds*. In fact, in illustration-style arguments the nature of scientific kinds shows up only as an afterthought — it is not our starting point or our main explanatory target (cf. Richmond, 2025a, 2025b, and recall the logic described in §3). This is in contrast to the way kinds are often approached by philosophers, who tend to focus on the nature of scientific kinds, assuming that this can tell us how the explanations invoking them work. This is currently the dominant approach whether the kind in question is *representation* (Shea, 2018), *function* (Garson, 2019), *gene* (Griffiths & Stotz, 2008), *computation* (Shagrir, 2022), or something else. The way I've approached representational explanation reverses the order of operations: we study a form of explanation and the resources it depends on, and conclusions about kinds, if they are necessary and relevant, fall out of that investigation.

Second, there is a whole bestiary of creatures that can be revisited in light of the distinction between example and illustration. As long as they are used carefully, thought experiments have the same legitimate role to play in philosophy of science that ablation experiments and

¹³ There are also implications for the literature on thought experiments, aside from the development of a more fine-grained logic to complement Häggqvist's (2009), as I noted in fn. 6. Thought experiments are often understood as a way to remove a system's features, usually to draw conclusions about *kind-hood* (Gendler, 2000): is it an X without the feature? As I describe in the main text, the conclusions to be drawn from Swampman are at best only secondarily about kind-hood; they are, first and foremost, about *how a system (scientific explanation) works*, in the same way that anuran vision experiments support conclusions about how the frog visual system works. My account is thus in line with philosophers who argue that thought experiments (typically in science, but we can extend their view to thought experiments in philosophy) should be understood on the model of real experiments (Gooding, 1992; Häggqvist, 2009; Schabas et al., 2018; Sorensen, 1992; Stuart, 2016). With feature-removal understood as an *experimental manipulation*, Swampman becomes an experimental stimulus akin to a cardboard worm. This also puts me in agreement with philosophers who stress the cognitive aspects of thought experiments, including imagination, the use of tacit or background knowledge, and the manipulation of mental models (el Skaf & Stuart, 2024; Mišević, 1992; Nersessian, 1992). In addition to those cognitive processes, I stress the *reasoning* we undertake when performing a thought experiment. This reasoning includes some processes (like experimental manipulation and prediction) that might be reducible to induction, consistent with authors who view thought experiments as arguments (Brendel, 2018, p. 291; Norton, 1996, 2004). But it seems to also include reasoning processes that don't reduce straightforwardly to argumentation, like the use of embodied conceptual knowledge, know-how, and narratives (Nersessian, 2018) to elaborate and constrain our understanding of the fictional situation, and to enable predictions about the two labs' responses to Swampman and Davidson.

environmental manipulations do in neuroscience and psychology. This is not to say the whole bestiary is welcome! Philosophical zombies, e.g., might not have much to say about consciousness science. But we should remain open to thought experiments as I've described them, with specific roles to play in establishing conclusions about scientific reasoning. We should dismiss them if their details undermine their role in drawing those conclusions; we should not dismiss them simply because they are unrealistic. Teleosemanticists in particular should appreciate this, since, to my knowledge at least, they have not attempted to defend the creatures they take seriously, like Kimus and Snorfs (Pietroski, 1992), from Millikan's criticisms, nor shown how those creatures can be informative about representation while Swampman is not.

For philosophy of cognitive science, there are further implications. First, there is a new Swampman sighting for teleosemanticists to debunk — and this one, I think, is much clearer about Swampman and his nature than the blurry and partial photos that previous expeditions have come back with. If teleosemanticists are right, they should be able to show what the Swampman illustration gets wrong about cognitive scientific practice. To do this, it is not enough to build a notion of representation that *could* serve scientific practice (Shea, 2018); it also means showing that this notion *is* used in scientific practice, and that means showing how the relevant explanations would fail in cases, like Swampman, that lack the features teleosemanticists think representational explanations rely on.

Second, and finally, the Swampman argument seems to show that some things in biology make quite a bit of sense *without* the light of evolution (to mangle the famous quote from Dobzhansky, 1973). Arguments for teleosemantics are often prefaced with remarks about how important it is to take evolution seriously (e.g., Garson, 2019, Chapter 12), and it's worth asking how we can do that if selection does not define cognitive scientific kinds like *representation*. Evolution doesn't become irrelevant to cognitive science, of course. But the murky history of the brain is only weakly informative about its current structure (de Sousa et al., 2023), and if evolutionary considerations don't define kinds like *representation*, their role in cognitive science has to be illuminated by examining the way cognitive scientific explanations use (or should use) those considerations — partly, as in the arguments above, by seeing how those considerations affect the success of different forms of explanation. That's an exciting project (cf. Cisek & Hayden, 2022), and hopefully one that reformed teleosemanticists will take up.

Acknowledgements: Thanks to audiences at Columbia University, Western University, and the Canadian Philosophical Association for feedback, and thanks to André Curtis-Trudel, Linus Huang, John Morrison, Kate Pendoley, Jason Winning, and the EMRG Lab for helpful discussions. And thanks to some anonymous reviewers for helpful comments.

Funding information: None to declare.

Conflicts: None to declare.

References

- Antony, L. (1996). Equal Rights for Swamp-persons. *Mind & Language*, 11(1), 70–75.
- Baggs, E., & Sanches de Oliveira, G. (2024). Rewilding psychology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1910), 20230287.
<https://doi.org/10.1098/rstb.2023.0287>
- Baker, B., Lange, R. D., Richmond, A., Kriegeskorte, N., Cao, R., Pitkow, X., Schwartz, O., & Achille, A. (n.d.). Use and usability: Concepts of representation in philosophy, neuroscience, cognitive science, and computer science. *Neurons, Behavior, Data, and Theory (Forthcoming)*.
- Barsalou, L. W. (2016). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review*, 23, 1122–1142. <https://doi.org/10.3758/s13423-016-1028-3>
- Bartolucci, G., Busiello, D. M., Ciarchi, M., Corticelli, A., Terlizzi, I. D., Olmeda, F., Revignas, D., & Schimmenti, V. M. (2025). *Phase behavior of Cacio and Pepe sauce* (No. arXiv:2501.00536). arXiv. <https://doi.org/10.48550/arXiv.2501.00536>
- Bickle, J., Mandik, P., & Landreth, A. (2019). The Philosophy of Neuroscience. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). Stanford University.
<https://plato.stanford.edu/archives/fall2019/entries/neuroscience/>
- Block, N. (1978). Troubles with Functionalism. In C. W. Savage (Ed.), *Perception and Cognition* (pp. 261–325). University of Minnesota Press.
- Boorse, C. (1976). Wright on Functions. *The Philosophical Review*, 85(1), 70–86.
- Brendel, E. (2018). The argument view: Are thought experiments mere picturesque arguments? In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge Companion to Thought Experiments* (pp. 281–292). Routledge.
- Cartwright, N. (2010). Models: Parables v Fables. In R. Frigg & M. Hunter (Eds.), *Beyond Mimesis and Convention* (Vol. 262, pp. 19–31). Springer Netherlands.
https://doi.org/10.1007/978-90-481-3851-7_2
- Chalmers, David. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Cisek, P., & Hayden, B. Y. (2022). Neuroscience needs evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1844).
<https://doi.org/10.1098/rstb.2020.0518>

- Cohnitz, D., & Häggqvist, S. (2018). Thought Experiments in Current Metaphilosophical Debates. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge Companion to Thought Experiments*. Routledge.
- Damásio, H., Grabowski, T., Frank, R., Galaburda, A. M., & Damásio, A. R. (1994). The return of phineas gage: Clues about the brain from the skull of a famous patient. *Science*, 264(5162), 1102–1105. <https://doi.org/10.4324/9780203496190>
- Davidson, D. (1987). Knowing One's Own Mind. *Proceedings and Addresses of the American Philosophical Association*, 441–458.
- de Sousa, A. A., Beaudet, A., Calvey, T., Bardo, A., Benoit, J., Charvet, C. J., Dehay, C., Gómez-Robles, A., Gunz, P., Heuer, K., van den Heuvel, M. P., Hurst, S., Lauters, P., Reed, D., Salagnon, M., Sherwood, C. C., Ströckens, F., Tawane, M., Todorov, O. S., ... Wei, Y. (2023). From fossils to mind. *Communications Biology*, 6(1), 1–21. <https://doi.org/10.1038/s42003-023-04803-4>
- Dennett, D. C. (1988). Precis of The Intentional Stance. *Behavioral and Brain Sciences*, 11, 495–546.
- Dennett, D. C. (2007). Cow-sharks, Magnets, and Swampman. *Mind & Language*.
- Descartes, R. (1984). Meditations on First Philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Eds.), *The Philosophical Writings of Descartes, Volume 2* (pp. 1–62). Cambridge University Press.
- Dobzhansky, T. (1973). Nothing in Biology Makes Sense except in the Light of Evolution. *The American Biology Teacher*, 35(3), 125–129.
- Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170, 115–135.
- Egan, F. (2022). The elusive role of normal-proper function in cognitive science. *Philosophy and Phenomenological Research*, 105(2), 468–475. <https://doi.org/10.1111/phpr.12930>
- el Skaf, R., & Stuart, M. T. (2024). Scientific Models and Thought Experiments: Same but Different. In R. Koskinen, N. Carrillo, & T. Knuuttila (Eds.), *The Routledge Handbook of Philosophy of Scientific Modeling*. Routledge.
- Garson, J. (2019). *What Biological Functions are and Why They Matter*. Cambridge University Press.
- Garson, J. (2022). Précis of Karen Neander's A Mark of the Mental . *Philosophy and Phenomenological Research*, 105(2), 461–467. <https://doi.org/10.1111/phpr.12933>

- Garson, J., & Papineau, D. (2019). Teleosemantics, selection and novel contents. *Biology and Philosophy*, 34(3), 1–20. <https://doi.org/10.1007/s10539-019-9689-8>
- Gendler, T. S. (2000). *Thought Experiment: On the Powers and Limits of Imaginary Cases*. Garland Publishing, Inc.
- Gibson, J. J. (1972). A theory of direct visual perception. In J. R. Royce & W. W. Rozeboom (Eds.), *The Psychology of Knowing* (pp. 215–240). Gordon & Breach.
- Goff, P., Seager, W., & Allen-Hermanson, S. (2022). Panpsychism. In *The Stanford Encyclopedia of Philosophy*.
- Gooding, D. C. (1992). What is Experimental about Thought Experiments? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1992(2), 280–290. <https://doi.org/10.1086/psaprocbienmeetp.1992.2.192842>
- Griffiths, P. E., & Stotz, K. (2008). Experimental philosophy of science. *Philosophy Compass*, 3(3), 507–521.
- Hacohen, O. (2022). The problem with appealing to history in defining neural representations. *European Journal for Philosophy of Science*, 12(3), 45. <https://doi.org/10.1007/s13194-022-00473-x>
- Häggqvist, S. (2009). A Model for Thought Experiments. *Canadian Journal of Philosophy*, 39(1), 55–76. <https://doi.org/10.1353/cjp.0.0040>
- Harlow, H. F. (1959). Love in Infant Monkeys. *Scientific American*, 200(6), 68–74. <https://doi.org/10.1038/scientificamerican0659-68>
- Hartle, B., & Wilcox, L. M. (2016). Depth magnitude from stereopsis: Assessment techniques and the role of experience. *Vision Research*, 125, 64–75. <https://doi.org/10.1016/j.visres.2016.05.006>
- Kim, D. (2021). Explanation and modality: On why the Swampman is still worrisome to teleosemanticists. *Synthese*, 199(1–2), 2817–2839. <https://doi.org/10.1007/s11229-020-02913-8>
- Lettvin, J. V., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the Frog's Eye Tells the Frog's Brain. *Proceedings of IRE*, 1940–1951.
- Lillian, P. E., Meyes, R., & Meisen, T. (2018). Ablation of a Robot's Brain: Neural Networks Under a Knife. *31st Conference on Neural Information Processing Systems*.

- Liu, F., Dai, S., Feng, D., Peng, X., Qin, Z., Kearns, A. C., Huang, W., Chen, Y., Ergün, S., Wang, H., Rappaport, J., Bryda, E. C., Chandrasekhar, A., Aktas, B., Hu, H., Chang, S. L., Gao, B., & Qin, X. (2019). Versatile cell ablation tools and their applications to study loss of cell functions. *Cellular and Molecular Life Sciences*, 76(23), 4725–4743. <https://doi.org/10.1007/s00018-019-03243-w>
- Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic Bulletin & Review*, 23, 941–958. <https://doi.org/10.3758/s13423-016-1045-2>
- Mekik, C. S., & Galang, C. M. (2022). Cognitive Science in a Nutshell. *Cognitive Science*, 46(8). <https://doi.org/10.1111/cogs.13179>
- Milkowski, M. (2013). *Explaining the Computational Mind*. MIT Press.
- Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press.
- Millikan, R. G. (1996). On Swampkinds. *Mind & Language*, 11(1), 103–117.
- Millikan, R. G. (2010). On Knowing the Meaning ; With a Coda on Swampman. *Mind*, 119(473), 43–81. <https://doi.org/10.1093/mind/fzpl57>
- Miščević, N. (1992). Mental models and thought experiments. *International Studies in the Philosophy of Science*. <https://doi.org/10.1080/02698599208573432>
- Mook, D. G. (1983). In Defense of External Invalidity. *American Psychologist*, 379–387.
- Nanay, B. (2014). Teleosemantics without etiology. *Philosophy of Science*, 81(5), 798–810. <https://doi.org/10.1086/677684>
- Neander, K. (1991). The teleological notion of “function.” *Australasian Journal of Philosophy*, 69(4), 454–468.
- Neander, K. (1996). Swampman Meets Swampcow. *Mind and Language*, 11(1), 118–129. <https://doi.org/10.1111/j.1468-0017.1996.tb00036.x>
- Neander, K. (2017). *A Mark of the Mental*. MIT Press.
- Nersessian, N. J. (1992). In the Theoretician’s Laboratory: Thought Experimenting as Mental Modeling. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 291–301.
- Nersessian, N. J. (2018). Cognitive science, mental modeling, and thought experiments. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge Companion to Thought Experiments* (pp. 309–326). Routledge.

- Norton, J. D. (1996). Are thought experiments just what you thought? *Canadian Journal of Philosophy*, 26, 333–366.
- Norton, J. D. (2004). On thought experiments: Is there more to the argument? *Philosophy of Science*, 71, 1139–1151.
- Papineau, D. (2001). The status of teleosemantics, or how to stop worrying about swampman. *Australasian Journal of Philosophy*, 79(2), 279–289. <https://doi.org/10.1080/713659227>
- Paul, L. A. (2014). *Transformative Experience*. Oxford University Press.
- Peacocke, C. (2014). Perception, biology, action, and knowledge. *Philosophy and Phenomenological Research*, 88(2), 477–484. <https://doi.org/10.1111/phpr.12092>
- Peters, U. (2014). Teleosemantics, Swampman, and Strong Representationalism. *Grazer Philosophische Studien*, 90, 273–288.
- Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press.
- Pietroski, P. M. (1992). Intentionality and teleological error. *Pacific Philosophical Quarterly*, 73(3), 267–282.
- Porter, B. (2020). Teleosemantics and tetrachromacy. *Biology and Philosophy*, 35(1), 1–22. <https://doi.org/10.1007/s10539-019-9732-9>
- Putnam, H. (1992). Brains in a Vat. In K. DeRose & T. A. Warfield (Eds.), *Skepticism: A Contemporary Reader* (pp. 187–204). Oxford University Press.
- Richmond, A. (2025a). How computation explains. *Mind & Language*, 40(1), 2–20. <https://doi.org/10.1111/mila.12521>
- Richmond, A. (2025b). What is a theory of neural representation for? *Synthese*, 205(14).
- Salvalaggio, A., de Filippo De Grazia, M., Zorzi, M., de Schotten, M. T., & Corbetta, M. (2020). Post-stroke deficit prediction from lesion and indirect structural and functional disconnection. *Brain*, 143(7), 2173–2188. <https://doi.org/10.1093/brain/awaa156>
- Sartori, L. (2023). Putting the ‘Experiment’ back into the ‘Thought Experiment.’ *Synthese*, 201(2), 34. <https://doi.org/10.1007/s11229-022-04011-3>
- Schabas, M., Stuart, M. T., Fehige, Y., & Brown, J. R. (2018). Thought Experiments in Economics. In *The Routledge Companion to Thought Experiments*. Routledge.
- Schulte, P. (2020). Why mental content is not like water: Reconsidering the reductive claims of teleosemantics. *Synthese*, 197(5), 2271–2290. <https://doi.org/10.1007/s11229-018-1808-6>

- Searle, J. R. (1980). Minds , brains , and programs. *The Behavioral and Brain Sciences*, 3, 417–457.
- Sebastián, M. Á. (2017). Functions and mental representation: The theoretical role of representations and its real nature. *Phenomenology and the Cognitive Sciences*, 16(2), 317–336. <https://doi.org/10.1007/s11097-015-9452-9>
- Shagrir, O. (2022). *The Nature of Physical Computation*. Oxford University Press.
- Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.
- Shepard, R. N. (1984). Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking, and Dreaming. *Psychological Review*, 91(4), 417–447. <https://doi.org/10.1093/mind/xxv.3.415-b>
- Sorensen, R. A. (1992). *Thought Experiments*. Oxford University Press.
- Stuart, M. T. (2016). Norton and the Logic of Thought Experiments. *Axiomathes*, 26(4), 451–466. <https://doi.org/10.1007/s10516-016-9306-2>
- Tolly, J. (2021). Swampman: A dilemma for proper functionalism. *Synthese*, 198(7), 1725–1750. <https://doi.org/10.1007/s11229-018-1684-0>
- Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461. <https://doi.org/10.1038/nrn.2016.44>
- von Eckardt Klein, B. (1977). Inferring Functional Localization from Neurological Evidence. In E. Walker (Ed.), *Explorations in the biology of language* (pp. 27–66). MIT Press.
- Walsh, D. M., & Ariew, A. (1996). A Taxonomy of Functions. *Canadian Journal of Philosophy*, 26, 493–514.
- Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation. *Brain*, 97(1), 709–728. <https://doi.org/10.1093/brain/97.1.709>
- Weissman-Fogel, I., & Granovsky, Y. (2019). The virtual lesion approach to transcranial magnetic stimulation: Studying the brain-behavioral relationships in experimental pain. *Pain Reports*, 4(4), 1–12. <https://doi.org/10.1097/PR9.0000000000000760>
- Wright, L. (1973). Functions. *Philosophical Review*, 82, 139–168.

Zhang, C., Kaye, J. A., Cai, Z., Wang, Y., Prescott, S. L., & Liberles, S. D. (2021). Area Postrema Cell Types that Mediate Nausea-Associated Behaviors. *Neuron*, 109(3), 461-472.e5. <https://doi.org/10.1016/j.neuron.2020.11.010>