Philosophy of Science (2025), **00**, 1–21
doi:10.1017/psa.2025.10106

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

What Really Lives in the Swamp? Thought Experiments and the Illustration of Scientific Reasoning

Andrew Richmond

Rotman Institute of Philosophy, Western University, London, ON, Canada
Email: arichmo8@uwo.ca

(Received 13 September 2024; revised 02 July 2025; accepted 08 July 2025)

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 as *examples of scientific kinds*. The logic of thought experiments, understood as illustrations, is analogous to the logic of common experimental paradigms in science. So, in reviving Swampman and showing how he survives teleosemantic objections, 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[1]

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

[1] (He's joking.)

© The Author(s), 2025. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

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 that 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 (1984) killed Swampman dead pretty much the moment he rose from the swamp. 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 article develops an account of thought experiments to show how creatures like Swampman *can* be relevant to our understanding of science. In section 2, I review Swampman, along with Millikan's argument against him, which has become the teleosemanticist's stock argument. In section 3, I 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 show that illustrations don't rely on the assumptions Millikan criticizes, and I reinforce this by comparing the logic of illustrations with some experimental paradigms in cognitive science. In section 4, I show that Swampman can be fruitfully understood as an illustration. So understood, he is untouched by standard objections and has interesting things to teach us about cognitive science. I conclude, in section 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]

[2] I'm ignoring teleosemanticists who appeal to nonselectional functions, such as forward-looking functions (Nanay 2014). They are not Swampman's target, or mine.

[3] I'm setting aside teleosemanticists concerned solely with *nonscientific* 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.

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, because 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]

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 (1996) 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. (108)

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 (especially clear about this motivation are Garson 2022, 464; Neander 1996, 120; Shea 2018, 22, 28–29). 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, 28–29), is that our intuitions about which kinds Swampman belongs to should be overridden by facts

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

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 that 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, 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, 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 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 be understood not as

[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, that is, 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 and 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.

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 spells this out in detail.

3. Examples and illustrations

In section 3.1, I develop a simple thought experiment, analogous to Swampman, that doesn't make problematic assumptions about kind membership. I call this sort of thought experiment an *illustration*.[7] In section 3.2, I compare the logic of illustration with the logic of some experimental paradigms in cognitive science. Partly, this is building to the argument in section 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 were to persist ("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 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 the following. 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

[7] Not to be confused with the "illustrative thought experiments" of Cohnitz and 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.

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, Seager, and Allen-Hermanson 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*, then deriving the nature of the kind from that example. In the terminology I 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 nonconscious world, but we're not starting from a claim that the nonconscious world is a member of some kind, 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 were 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. 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 down the logic of illustration 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 were 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 and Galang 2022, 2), specifically how they use different resources to perform them—both internal resources (such as neural structures or activity patterns) and external ones (such as 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 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, for example, by a stroke or a railroad spike (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 and Hickok 2016, 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 and 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, Mandik, and Landreth 2019, 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.

[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., forthcoming).

[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 for which removing a feature has no effect on explanations, so I'll focus on the corresponding subset of ablation studies.

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?

I wanted to do two things 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 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, 709), and there are even attempts to draw conclusions about human brains from ablations in artificial neural networks (Lillian, Meyes, and Meisen 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, that is, 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, 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, for example, 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 succeeded even if they originally *did* rely on the

[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.

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 setup of that paper tells us that a frog "will starve to death surrounded by food if [the food] is not moving" (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" (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 (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 (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 nonnaturalistic 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 about kind membership. They would not say, "Well, these are very nice experiments, but unfortunately, you've made a terrible error. The nonnaturalistic 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 nonnaturalistic 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 (nonnaturalistic 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 nonnaturalistic environments. Ecological psychologists have been making points like this for decades. In impoverished (experimental) environments, organisms may perform tasks differently than they do normally, when they have more environmental information to use (Baggs and Sanches de Oliveira 2024; Gibson 1972; but see Shepard 1984). The task for the frog researcher is to show that the *potential* differences in how prey capture works are not *actual*.

For a different example, consider Hartle and 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 artifactual *monocular* cues in the experimental stimuli. So when Hartle and Wilcox found that depth perception was not affected by the removal of binocular cues, it was 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, nonexperimental settings, you need a plausible argument that the "paring" didn't induce the organism to perform its task in a new way, for example, 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 section 4.1, I 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 discussed earlier do. In section 4.2, I 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's 2022 rendition of the argument is similar to the one that follows 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. Because 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's and Davidson's memories, 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 am 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, chap. 5), prey capture (Neander 2017, chap. 5), and the communication of resource locations (Millikan 1984, chap. 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]

[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, 22). This would make things easy for the teleosemanticist, but it isn't a plausible characterization of cognitive science. Cognitive scientists aim to explain, for example, 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

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 cognitive science can or can't show how Swampman's physical organization supports capacities like navigation, memory, and prey capture—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 were to sit 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 in which 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 just as predictive of Swampman's and Davidson's behavior, just as revealing of its neural basis, just 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. Hacohen 2022).

Again, I'm not talking about *all* projects cognitive science might have. A cognitive scientist might explain how an organism evolved a certain brain organization, and

simply patterns of behavior, or as successes in a nonselectional sense (compare note 2, and, for selection-agnostic notions of tasks and task success, see Baker et al., forthcoming; Mekik and Galang 2022). So while selection history may be an explanatory resource, it is not a precondition for thinking scientifically about cognition. Louise Antony (1996, 72) 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. The question is what resources they would rely on and whether those resources would include selection history.

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 those in 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 that we should not define *observation* in terms of consciousness. Because there are no problematic assumptions here about kind membership (the logic of the argument is the logic described in section 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 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 earlier 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 because, 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.[12]

[12] But I comment below 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 as 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 were to define representation in nonselectional terms. For example, it is sometimes suggested—or, more accurately, stipulated—that cognitive science is in the business of generalizing over species kinds (Millikan 1996, 109) and maximizing the reliability (Millikan 1996, 108) or breadth (Neander 1996, 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, which aim not only to generalize but also to model, problem solve, explain, and so on, as I noted in section 2. And anyway, it is not at all obvious that nonselectional kinds would fail to support the necessary generalizations and inductions—something that is acknowledged even by teleosemanticists (Shea 2018, 22). So this appeal to broader goals is not a plausible objection unless teleosemanticists can (as I described at the end of section 4.1) argue for some conception of those goals and show how representational explanation 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 and 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 history, the remaining scraps (short-term selection) 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 an 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, 1996; Millikan 1996)? It's tempting to brush off things like Swampman with nothing more than, "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 (1983) puts it, one cannot just point out that an experimental setting is unrealistic and reject it as externally invalid.[13] 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" (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 section 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 and Ariew 1996). The problem with these cases is that they can be nitpicked 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 precursor of a *similar enough* kind for it to count as selected for? The advantage of Swampman is the same advantage cardboard cutouts 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 successful only because we made him up to be as similar to Davidson as possible, allowing us to take advantage of *preexisting* 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. 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

[13] 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, allow us to draw conclusions about the world.

isolate particular features and 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, 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 *because Swampman would be explained in representational terms, he represents*. As I discussed in section 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 because representational explanations of Swampman would be just as successful as they are for 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 a further step and conclude that the kind *representation* shouldn't be defined in terms of selection history.

The significance of this reframing is revealed when we return to the *real kinds* response and Swampman's distance from real representational systems. As I put Millikan's point earlier, (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, 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 roles 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.[14]

[14] 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 mentioned in note 6. Thought experiments are often understood as a way to remove a system's features, usually to draw conclusions

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 section 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 and 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; conclusions about kinds, if they are necessary and relevant, fall out of that investigation.

Second, a whole bestiary of creatures 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 environmental manipulations do in neuroscience and psychology. This is not to say that the whole bestiary is welcome! Philosophical zombies, for example, 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, because, to my knowledge at least, they have not attempted to defend the creatures they take seriously, such as Kimus and Snorfs (Pietroski 1992), from Millikan's criticisms—they have not 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 one that is much

about *kind-hood*: Is it an *X* without the feature (Gendler 2000)? 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 those of 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 and Stuart 2024; Miščević 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, 291; Norton 1996, 2004). But it seems to also include reasoning processes that don't reduce straightforwardly to argumentation, such as 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.

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, chap. 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 preceding arguments, by seeing how those considerations affect the success of different forms of explanation. That's an exciting project (cf. Cisek and Hayden 2022) and, we can hope, one that reformed teleosemanticists will take up.

Acknowledgments. 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.

References

Antony, Louise. 1996. "Equal Rights for Swamp-Persons." *Mind and Language* 11 (1):70–75. https://doi.org/10.1111/j.1468-0017.1996.tb00030.x.

Baggs, Edward, and Guilherme Sanches de Oliveira. 2024. "Rewilding Psychology." *Philosophical Transactions of the Royal Society B* 379 (1910):20230287. https://doi.org/10.1098/rstb.2023.0287.

Baker, Ben, Richard Lange, Andrew Richmond, Nikolaus Kriegeskorte, Rosa Cao, Xaq Pitkow, and Odelia Schwartz. Forthcoming. "Use and Usability: Concepts of Representation in Philosophy, Neuroscience, Cognitive Science, and Computer Science." *Neurons, Behavior, Data, and Theory*.

Barsalou, Lawrence W. 2016. "On Staying Grounded and Avoiding Quixotic Dead Ends." *Psychonomic Bulletin and Review* 23:1122–42. https://doi.org/10.3758/s13423-016-1028-3.

Bartolucci, Giacomo, Daniel Maria Busiello, Matteo Ciarchi, Alberto Corticelli, Ivan Di Terlizzi, Fabrizio Olmeda, Davide Revignas, and Vincenzo Maria Schimmenti. 2025. "Phase behavior of Cacio e Pepe sauce." *Physics of Fluids*, 37(4), 044122. https://doi.org/10.1063/5.0255841

Bickle, John, Peter Mandik, and Anthony Landreth. 2019. "The Philosophy of Neuroscience." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2019/entries/neuroscience/.

Block, Ned. 1978. "Troubles with Functionalism." In *Perception and Cognition*, edited by C. Wade Savage, 261–325. University of Minnesota Press.

Boorse, Christopher. 1976. "Wright on Functions." *Philosophical Review* 85 (1):70–86. https://doi.org/10.2307/2184255.

Brendel, Elke. 2018. "The Argument View: Are Thought Experiments Mere Picturesque Arguments?" In *The Routledge Companion to Thought Experiments*, edited by Michael T. Stuart, Yiftach Fehige, and James Robert Brown, 281–92. Routledge.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Cisek, Paul, and Benjamin Y. Hayden. 2022. "Neuroscience Needs Evolution." *Philosophical Transactions of the Royal Society B* 377 (1844):20200518. https://doi.org/10.1098/rstb.2020.0518.

Cohnitz, Daniel, and Sören Häggqvist. 2018. "Thought Experiments in Current Metaphilosophical Debates." In *The Routledge Companion to Thought Experiments*, edited by Michael T. Stuart, Yiftach Fehige, and James Robert Brown, 406–424. Routledge.

Damásio, Hanna, Thomas Grabowski, Randall Frank, Albert M. Galaburda, and Antonio R. Damasio. 1994. "The Return of Phineas Gage: Clues About the Brain from the Skull of a Famous Patient." *Science* 264 (5162):1102–5. https://doi.org/10.4324/9780203496190.

Davidson, Donald. 1987. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60 (3):441–58. https://doi.org/10.2307/3131782.

Dennett, Daniel C. 1988. "Precis of the Intentional Stance." *Behavioral and Brain Sciences* 11 (3):495–546.

Dennett, Daniel C. 1996. "Cow-Sharks, Magnets, and Swampman." *Mind and Language* 11 (1):76–77.

Descartes, René. 1984. "Meditations on First Philosophy." In *The Philosophical Writings of Descartes*, vol. 2, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch, 1–62. Cambridge University Press.

de Sousa, Alexandra A., Amélie Beaudet, Tanya Calvey, Ameline Bardo, Julien Benoit, Christine J. Charvet, Colette Dehay, et al. 2023. "From Fossils to Mind." *Communications Biology* 6 (1):1–21. https://doi.org/10.1038/s42003-023-04803-4.

Dobzhansky, Theodosius. 1973. "Nothing in Biology Makes Sense Except in the Light of Evolution." *American Biology Teacher* 35 (3):125–29. https://doi.org/10.2307/4444260.

Egan, Frances. 2014. "How to Think About Mental Content." *Philosophical Studies* 170:115–35. https://doi.org/10.1007/s11098-013-0172-0.

Egan, Frances. 2022. "The Elusive Role of Normal-Proper Function in Cognitive Science." *Philosophy and Phenomenological Research* 105 (2):468–75. https://doi.org/10.1111/phpr.12930.

el Skaf, Rawad, and Michael T. Stuart. 2024. "Scientific Models and Thought Experiments: Same but Different." In *The Routledge Handbook of Philosophy of Scientific Modeling*, edited by Rami Koskinen, Natalia Carrillo, and Tarja Knuuttila, 325–340. Routledge.

Garson, Justin. 2019. *What Biological Functions Are and Why They Matter*. Cambridge University Press.

Garson, Justin. 2022. "Précis of Karen Neander's 'A Mark of the Mental.'" *Philosophy and Phenomenological Research* 105 (2):461–67. https://doi.org/10.1111/phpr.12933.

Garson, Justin, and David Papineau. 2019. "Teleosemantics, Selection and Novel Contents." *Biology and Philosophy* 34 (3):1–20. https://doi.org/10.1007/s10539-019-9689-8.

Gendler, Tamar Szabo. 2000. *Thought Experiment: On the Powers and Limits of Imaginary Cases*. Garland.

Gibson, James. J. 1972. "A Theory of Direct Visual Perception." In *The Psychology of Knowing*, edited by Jospeh R. Royce and William W. Rozeboom, 215–40. Gordon and Breach.

Goff, Philip, William Seager, and Sean Allen-Hermanson. 2022. "Panpsychism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/panpsychism/.

Gooding, David C. 1992. "What Is Experimental About Thought Experiments?" *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992 (2):280–90. https://doi.org/10.1086/psaprocbienmeetp.1992.2.192842.

Griffiths, Paul E., and Karola Stotz. 2008. "Experimental Philosophy of Science." *Philosophy Compass* 3 (3):507–21. https://doi.org/10.1111/j.1747-9991.2008.00133.x.

Hacohen, Ori. 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, Sören. 2009. "A Model for Thought Experiments." *Canadian Journal of Philosophy* 39 (1):55–76. https://doi.org/10.1353/cjp.0.0040.

Harlow, Harry F. 1959. "Love in Infant Monkeys." *Scientific American* 200 (6):68–74. https://doi.org/10.1038/scientificamerican0659-68.

Hartle, Brittney, and Laurie M. Wilcox. 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, Dongwu. 2021. "Explanation and Modality: On Why the Swampman Is Still Worrisome to Teleosemanticists." *Synthese* 199 (1–2):2817–39. https://doi.org/10.1007/s11229-020-02913-8.

Lettvin, Jerome, Humberto Maturana, Warren McCulloch, and Walter H. Pitts. 1959. "What the Frog's Eye Tells the Frog's Brain." *Proceedings of IRE* 47 (11):1940–51. https://doi.org/10.1109/JRPROC.1959.287207.

Lillian, Peter E., Richard Meyes, and Tobias Meisen. 2018. "Ablation of a Robot's Brain: Neural Networks Under a Knife." Paper presented at the 31st Conference on Neural Information Processing Systems.

Liu, Fengming, Shen Dai, Dechun Feng, Xiao Peng, Zhongnan Qin, Alison C. Kearns, Wenfei Huang, et al. 2019. "Versatile Cell Ablation Tools and Their Applications to Study Loss of Cell Functions." *Cellular and Molecular Life Sciences* 76 (23):4725–43. https://doi.org/10.1007/s00018-019-03243-w.

Mahon, Bradford Z., and Gregory Hickok. 2016. "Arguments About the Nature of Concepts: Symbols, Embodiment, and Beyond." *Psychonomic Bulletin and Review* 23:941–58. https://doi.org/10.3758/s13423-016-1045-2.

Mekik, Can S., and Carl Michael Galang. 2022. "Cognitive Science in a Nutshell." *Cognitive Science* 46 (8): e13179. https://doi.org/10.1111/cogs.13179.

Milkowski, Marcin. 2013. *Explaining the Computational Mind*. MIT Press.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories*. MIT Press.

Millikan, Ruth Garrett. 1996. "On Swampkinds." *Mind and Language* 11 (1):103–17.

Millikan, Ruth Garrett. 2010. "On Knowing the Meaning; with a Coda on Swampman." *Mind* 119 (473):43–81. https://doi.org/10.1093/mind/fzpl57.

Miščević, Nenad. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6 (3):215–26. https://doi.org/10.1080/02698599208573432.

Mook, Douglas G. 1983. "In Defense of External Invalidity." *American Psychologist* 38 (4):379–87. https://doi.org/10.1037/0003-066X.38.4.379.

Nanay, Bence. 2014. "Teleosemantics Without Etiology." *Philosophy of Science* 81 (5):798–810. https://doi.org/10.1086/677684.

Neander, Karen. 1991. "The Teleological Notion of 'Function.'" *Australasian Journal of Philosophy* 69 (4):454–68. https://doi.org/10.1080/00048409112344881.

Neander, Karen. 1996. "Swampman Meets Swampcow." *Mind and Language* 11 (1):118–29. https://doi.org/10.1111/j.1468-0017.1996.tb00036.x.

Neander, Karen. 2017. *A Mark of the Mental*. MIT Press.

Nersessian, Nancy J. 1992. "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling." *Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992 (2):291–301. https://doi.org/10.1086/psaprocbienmeetp.1992.2.192843.

Nersessian, Nancy J. 2018. "Cognitive Science, Mental Modeling, and Thought Experiments." In *The Routledge Companion to Thought Experiments*, edited by Michael T. Stuart, Yiftach Fehige, and James Robert Brown, 309–26. Routledge.

Norton, John D. 1996. "Are Thought Experiments Just What You Thought?" *Canadian Journal of Philosophy* 26 (3):333–66. https://doi.org/10.1080/00455091.1996.10717457.

Norton, John D. 2004. "On Thought Experiments: Is There More to the Argument?" *Philosophy of Science* 71 (5):1139–51. https://doi.org/10.1086/425238.

Papineau, David. 2001. "The Status of Teleosemantics, or How to Stop Worrying About Swampman." *Australasian Journal of Philosophy* 79 (2):279–89. https://doi.org/10.1080/713659227.

Paul, Laurie. A. 2014. *Transformative Experience*. Oxford University Press.

Peacocke, Christopher. 2014. "Perception, Biology, Action, and Knowledge." *Philosophy and Phenomenological Research* 99 (2):477–84. https://doi.org/10.1111/phpr.12092.

Peters, Uwe. 2014. "Teleosemantics, Swampman, and Strong Representationalism." *Grazer Philosophische Studien* 90 (1):273–88. https://doi.org/10.1163/9789004298767_017.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanistic Account*. Oxford University Press.

Pietroski, Paul M. 1992. "Intentionality and Teleological Error." *Pacific Philosophical Quarterly* 73 (3):267–82. https://doi.org/10.1111/j.1468-0114.1992.tb00339.x.

Porter, Brian. 2020. "Teleosemantics and Tetrachromacy." *Biology and Philosophy* 35 (1):1–22. https://doi.org/10.1007/s10539-019-9732-9.

Putnam, Hilary. 1992. "Brains in a Vat." In *Skepticism: A Contemporary Reader*, edited by Keith DeRose and Ted A. Warfield, 187–204. Oxford University Press.

Richmond, Andrew. 2025a. "How Computation Explains." *Mind and Language* 40 (1):2–20. https://doi.org/10.1111/mila.12521.

Richmond, Andrew. 2025b. "What Is a Theory of Neural Representation For?" *Synthese* 205:14. https://doi.org/10.1007/s11229-024-04816-4.

Salvalaggio, Alessandro, Michele de Filippo De Grazia, Marco Zorzi, Michel Thiebaut de Schotten, and Maurizio Corbetta. 2020. "Post-Stroke Deficit Prediction from Lesion and Indirect Structural and Functional Disconnection." *Brain* 143 (7):2173–88. https://doi.org/10.1093/brain/awaa156.

Sartori, Lorenzo. 2023. "Putting the 'Experiment' Back into the 'Thought Experiment.'" *Synthese* 201 (2):34. https://doi.org/10.1007/s11229-022-04011-3.

Schabas, Margaret. 2018. "Thought Experiments in Economics." In *The Routledge Companion to Thought Experiments*, edited by Michael T. Stuart, Yiftach Fehige, and James Robert Brown, 171–182. Routledge.

Schulte, Peter. 2020. "Why Mental Content Is Not Like Water: Reconsidering the Reductive Claims of Teleosemantics." *Synthese* 197 (5):2271–90. https://doi.org/10.1007/s11229-018-1808-6.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3):417–57. https://doi.org/10.1017/S0140525X00005756.

Sebastián, Miguel Ángel. 2017. "Functions and Mental Representation: The Theoretical Role of Representations and Its Real Nature." *Phenomenology and the Cognitive Sciences* 16 (2):317–36. https://doi.org/10.1007/s11097-015-9452-9.

Shagrir, Oron. 2022. *The Nature of Physical Computation*. Oxford University Press.

Shea, Nicholas. 2018. *Representation in Cognitive Science*. Oxford University Press.

Shepard, Roger. N. 1984. "Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking, and Dreaming." *Psychological Review* 91 (4):417–47. https://doi.org/10.1093/mind/xxv.3.415-b.

Sorensen, Roy A. 1992. *Thought Experiments*. Oxford University Press.

Stuart, Michael T. 2016. "Norton and the Logic of Thought Experiments." *Axiomathes* 26 (4):451–66. https://doi.org/10.1007/s10516-016-9306-2.

Tolly, Jeffrey. 2021. "Swampman: A Dilemma for Proper Functionalism." *Synthese* 198 (7):1725–50. https://doi.org/10.1007/s11229-018-1684-0.

Tononi, Guilio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. "Integrated Information Theory: From Consciousness to Its Physical Substrate." *Nature Reviews Neuroscience* 17 (7):450–61. https://doi.org/10.1038/nrn.2016.44.

von Eckardt Klein, Barbara. 1977. "Inferring Functional Localization from Neurological Evidence." In *Explorations in the Biology of Language*, edited by Edward Walker, 27–66. MIT Press.

Walsh, Denis M., and André Ariew. 1996. "A Taxonomy of Functions." *Canadian Journal of Philosophy* 26 (4):493–514. https://doi.org/10.1080/00455091.1996.10717464.

Waters, Kenneth. 2019. An Epistemology of Scientific Practice. *Philosophy of Science*, 86 (4):585–611. https://doi.org/10.1086/704973.

Weiskrantz, Larry, Elizabeth K. Warrington, Michael D. Sanders, and John Marshall. 1974. "Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation." *Brain* 97 (1):709–28. https://doi.org/10.1093/brain/97.1.709.

Weissman-Fogel, Irit, and Yelena Granovsky. 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, Larry. 1973. "Functions." *Philosophical Review* 82 (2):139–68. https://doi.org/10.2307/2183766.

Zhang, Chuchu, Judith A. Kaye, Zerong Cai, Yandan Wang, Sara L. Prescott, and Stephen D. Liberles. 2021. "Area Postrema Cell Types That Mediate Nausea-Associated Behaviors." *Neuron* 109 (3):461–72. https://doi.org/10.1016/j.neuron.2020.11.010.

Cite this article: Richmond, Andrew. 2025. "What Really Lives in the Swamp? Thought Experiments and the Illustration of Scientific Reasoning." *Philosophy of Science*. https://doi.org/10.1017/psa.2025.10106