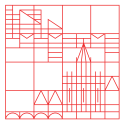
WHATIF
WASWÄREWENN

ON THE MEANING, RELEVANCE, AND EPISTEMOLOGY OF COUNTERFACTUAL CLAIMS AND THOUGHT EXPERIMENTS

Universität Konstanz



Brian Leahy & Maximilian Huber

TWO ARGUMENTS FOR THE ETIOLOGICAL THEORY OVER THE MODAL THEORY OF FUNCTION

UNIVERSITY OF KONSTANZ ◆ PREPRINT SERIES ◆ NO 3 / 2014

WHATIF
WASWÄREWENN

2013

ISSUE #	AUTHOR NAME	TITLE
1	WOLFGANG SPOHN	A RANKING-THEORETIC APPROACH TO CONDITIONALS
2	WOLFGANG SPOHN	AGM, RANKING THEORY, AND THE MANY WAYS TO COPE WITH EXAMPLES
3	MARCEL WEBER	CAUSAL SELECTION VS CAUSAL PARITY IN BIOLOGY: RELEVANT COUNTERFACTUALS AND BIOLOGICALLY NORMAL INTERVENTIONS
4	AMAIA GARCIA-ODON	PROJECTION AND CONDITIONALIZATION

WHATIF
WASWÄREWENN

2014

ISSUE #	AUTHOR NAME	TITLE
1	NATASHA GRIGORIAN	THOMAS MALTHUS AND NIKOLAI CHERNYSHEVSKY: THOUGHT EXPERIMENTS AND VISIONS OF THE FUTURE
2	DANIEL DOHRN	BENNETT WORLDS AND THE ASYMMETRY OF COUNTERFACTUAL DEPENDENCE
3	BRIAN LEAHY & MAXIMILIAN HUBER	TWO ARGUMENTS FOR THE ETIOLOGICAL THEORY OVER THE MODAL THEORY OF FUNCTION

Two arguments for the etiological theory over the modal theory of function*

Brian Leahy
University of Konstanz

Maximilian Huber
University of Geneva

Abstract This paper contains a positive development and a negative argument. It develops a theory of function loss and shows how this undermines an objection raised against the etiological theory of function in support of the modal theory of function. Then it raises two internal problems for the modal theory of function.

Keywords: Biological function, modal theory, counterfactual implication, etiological theory, David Lewis, Ruth Millikan

1 Introduction

It is theoretically useful to ascribe functions not only to artifacts like corkscrews and tape measures, which have been designed to perform some function, but also to biological traits like hearts and immune systems. Functions of biological traits, however, cannot be analyzed in terms of the intentions of their designers without invoking the existence of a designer. Several nonintentional theories of function have arisen as a consequence.

Nanay (2010) argues that there is no coherent, noncircular way to analyze the functions of a trait token in terms of properties of other tokens of the same trait type. Any such analysis requires a method for individuating trait types, that is, for determining whether or not two trait tokens are members of the same type. He considers three criteria for determining when two trait tokens are members of the same type, and argues that all are inadequate. This is called *the problem of trait type individuation*. Nanay then offers the *modal theory of function*, which analyzes the functions of a trait token in terms of the properties of that trait token itself. The functional properties of a token trait at a time are determined by which counterfactuals are true of that trait at that time.

The principal selling point of the modal theory is that it avoids the problem of trait type individuation. We offer two arguments against the modal theory. First, we

* We would like to thank Bence Nanay, Johannes Schmitt, John Troyer, and the members of the Lake Geneva Biological Interest Group for their valuable input on earlier versions of this paper. B. Leahy was supported by DFG Research Group 1614 "What if: On the meaning, relevance, and epistemology of counterfactual claims and thought experiments". M. Huber was supported by the Swiss National Science Foundation grant 140885.

construct a theory of function loss that resolves the problem of trait type individuation for Millikan's etiological theory of function, thereby undermining the modal theory's main virtue. Second, we show that the modal theory has two internal problems. First, any theory of function must be able to distinguish instances of function performance from behaviours that increase fitness by accident. Attention to the details of available formal semantics for counterfactuals shows that the modal theory is unable to draw this distinction. Second, any theory of function must account for malfunction and dispositional functions in off-mode. But in order to account for malfunction and dispositional functions in off-mode the modal theory must either adopt a theory of transworld identity or fall subject to the problem of trait type individuation.

2 A Theory of Function Loss

2.1 Nanay's Challenge

Very roughly, Millikan's (1984) etiological theory maintains that a token trait x of organism O has function F iff F-ing is the effect that some ancestor token x' of x was selected for. Trait types on this analysis are individuated in terms of homology (hence the appeal to ancestry). Nanay argues that trait types cannot be adequately individuated in terms of homology, morphology, or function. We briefly mention the problems for function and morphology before turning to the problem for homology.

Function. A trait token x belongs to trait type X iff x has some function F. For example, a token trait is a heart if and only if it has the function to pump blood. Nanay (2010: 415f.) claims that the functional criterion is untenable on pain of vicious circularity.[1]

Morphology. A trait token x belongs to trait type X if and only if x has certain morphological properties. Nanay (2010: 416) rejects the morphological criterion because function is independent of morphology: a heart, no matter how malformed, still has the function of pumping blood. A virtue of the etiological theory of function is that it explains why individuals who are unable to perform functions due to departures from the morphological norm still have those functions; to adopt morphological criteria would be to give up this virtue.

Homology. Two traits tokens x and x' belong to the same trait type X iff x and x' are homologous. Two trait tokens x and x' are homologous iff x and x' have a common ancestral trait x''. For example, compare the wing of an eagle and the

1 This is contested, first, by Kiritani (2011) who argues that using different notions of function on the left and right hand side of the biconditional avoids the charge of vicious circularity; Nanay (2011) objects that this proposal fails to satisfy his desideratum of accounting for malfunctioning token traits. Second, Neander & Rosenberg (2012) argue that a reformulation of the etiological theory eliminates the presupposition of a criterion of trait-type individuation; Nanay (2012) disagrees.

forelimb of the first ancient tetrapod to crawl from the sea. The eagle's wings and the ancient tetrapod's forelimbs are homologous because the ancient tetrapod's forelimb is an ancestor of the eagle's wing. On this view, the eagle wing and ancient tetrapod forelimb are members of the same trait type. Nanay (2010: 418) objects that this "broad way of typing traits does not help in our definition of function, as using this way of talking about trait types would attribute the function of crawling to the eagle's wings". For the eagle's wing is a descendent of the tetrapod forelimb, which had the function of crawling. As we stated it, the etiological theory maintains that a particular eagle's wing has the function of crawling iff crawling is the effect that some ancestor of the wing was selected for. And this condition is met, since some ancient tetrapod had a forelimb that (a) is an ancestor of the eagle's wing and (b) was selected for crawling. Thus typing traits via homology yields unwanted function ascriptions for the etiological theory. In the following, we will use 'ancient tetrapod' to refer to individuals that (a) are ancestors of modern eagles and (b) have forelimbs with the crawling function.

Nanay then demonstrates the inadequacy of a possible response; the reason why this possible response is inadequate is informative so we review and clarify it here. The response maintains that two token traits x and x' belong to the same trait type iff x and x' are recent homologues. In other words, x and x' belong to the same trait type iff they have a common ancestral trait that does not date back farther than a certain threshold of years.

The response avoids attributing to eagle wings the crawling function because the most recent common ancestral trait between eagle wings and ancient tetrapod forelimbs dates too far back. However, it would struggle to explain why tokens of eagle and ancient tetrapod eyes belong to a single trait type with with a visual representation function. What could justify treating ancient tetrapod and eagle eyes as a single trait type while treating their forelimbs as different trait types? We believe that Nanay correctly identifies the difference:

> The only thing that differentiates the example of the eye from the example of the forelimb is that the selection pressure changed in the latter case, but did not change in the former. Forelimbs have been selected for doing something *different* in the bird population and in the ancient [tetrapod] population. Eyes, on the other hand, have been selected for doing *the same thing* in the bird population and in the ancient [tetrapod] population" (Nanay 2010: 418, emphasis original).

In Sect. 2.2 we exploit this difference to develop a theory of function loss that enables us to resolve Nanay's challenge while individuating trait types in terms of homology.

2.2 A Theory of Function Loss

The need for a theory of function loss is acknowledged in Millikan (1984: 32), though none is offered. In Sect. 2.2.1 we provide a theory of function loss with function replacement. This will be shown to adequately undermine Nanay's reasons for thinking that homology is inadequate for individuating trait types. In Sect. 2.2.2 we describe function loss without replacement. This constitutes a complete theory of function loss.

2.2.1 Function Replacement

Nanay challenges the etiological theory to explain (i) how eagle and ancient tetrapod forelimbs are both members of the same trait type, and (ii) why eagle forelimbs do not have a crawling function. And we must do so in a manner that allows us to say (iii) that eagle and ancient tetrapod eyes are both members of the same trait type, and (iv) that both have vision as function.

For Millikan (1984: 33f.), every function is associated with a normal explanation. A normal explanation is an explanation of how some function was most often performed, when it was performed. Two clarifications are in order. First, functions are sometimes performed by accident. Rabbit predator-detectors have as function to help conspecifics escape predators; this is most often accomplished by causing loud leg thumps when predators are observed. Skittish rabbits sometimes thump in response to non-predators such as falling branches. Sometimes rabbit thumps are caused by falling branches when undetected predators are in fact nearby; in such a case the function is performed, but not in accord with a normal explanation. That is why the definition of 'normal explanation' includes the qualifier 'most often'. Second, some functions are performed only rarely. Few sperm fertilize eggs. A normal explanation explains how sperm most often fertilize eggs when they manage to do so. This is why the definition includes the qualifier 'when it was performed'.

A normal explanation for the performance of a function will appeal to internal characteristics shared by all trait tokens that performed the function in accord with the normal explanation, environmental conditions that all trait tokens were in when performing the function in accord with the normal explanation and, by addition of natural laws, show how this arrangement guaranteed the performance of the function. For example, a normal explanation of the eagle's wing's flying function will describe internal characteristics such as the size and shape of eagle wings; the hollow bones that make the wings light; the shape and arrangement of feathers, etc; it will describe environmental conditions such as the density of air that contributes to flight and the mass of the Earth; it will add natural laws such as the relationship between mass and gravity and Bernoulli's principle all in sufficient detail to show how this arrangement

guarantees that the wing causes flight when all the other factors are in place. We will call the internal characteristics described by the normal explanation for a function *normal internal characteristics* for that function.

Ancient tetrapod forelimbs had a crawling function, and there is a normal explanation for the performance of that function that describes normal internal characteristics of ancient tetrapod forelimbs. The normal explanation for the tetrapod's crawling function requires not only different internal characteristics than those required for normal performance of the flying function, but incompatible characteristics. We propose that if selection pressures on a family of traits have changed so that the normal performance of a more recent function requires an internal characteristic that is incompatible with some internal characteristic required for normal performance of a less recent function, then members of the family that have the more recent function have lost the less recent function. The less recent function has been replaced by the more recent function.

We pause to elaborate on the notion of incompatible characteristics. Consider the normal internal characteristics of eagle wings and tetrapod forelimbs. Eagle wings enable flight in part due to having a shape that enables air molecules to spread over them in a particular way. But having that shape makes them ill suited to bearing much body weight. Ancient tetrapod forelimbs, on the other hand, enabled crawling in part because they were suited to bearing a large portion of the individual's body weight. In other words, there is a trade-off between avian locomotion and crawling on all fours. We do not know whether this trade-off is a matter of necessity. Perhaps there were (or are) wings that enable flight and double as weight-bearing forelimbs. But as a matter of fact, eagle wings have evolved in a manner such that their normal internal characteristics as required for flight make them unsuited to bearing weight, as required for crawling in accord with a normal explanation (cf. (Gatesy & Dial 1996: 333f.)).

Note that merely having characteristics incompatible with the performance of a function is not sufficient for function loss. A rabbit with hind legs too small to thump loudly has not thereby lost the function of warning conspecifics. For the small hind legs of this mutant rabbit are not normal internal characteristics required for normal performance of any proper function that is more recent than the conspecific warning function.

The condition provided so far is not necessary for function loss. There are other ways to lose functions, explored in Sect. 2.2.2. For example, consider a vestigial trait such as the human appendix: it has lost the function of digesting cellulose, however, this will not be delivered by our sufficient condition unless modern appendixes have a new function that requires normal internal characteristics incompatible with those required for digesting cellulose.

The proposal resolves the problem of trait type individuation. Eagle and ancient

tetrapod forelimbs are members of the same trait type in virtue of common descent. Eagle wings lack the crawling function because the more recent flying function has been selected for in virtue of internal characteristics that are incompatible with internal characteristics required by the less recent crawling function. Eagle and ancient tetrapod eyes are members of the same trait type in virtue of common descent. Both have vision as function because causing vision is what they were selected for and in neither case is there a more recent function that has been selected for in virtue of characteristics incompatible with those required for vision via the normal internal characteristics. So (i)-(iv) are satisfied.

The circularity that impeded functional criteria for trait type identification poses no threat here. Traits are typed in terms of their phylogeny. Functions of token traits are analyzed in terms of effects of earlier tokens of the same trait type. A sufficient condition for loss of a function F within a (sub)family is the attainment of a novel function F' whose normal performance requires internal characteristics incompatible with those required for normal performance of F.

2.2.2 Function loss without replacement

Perhaps traits can lose functions without acquiring any new function. Though not required for resolving the problem of trait type individuation, this possibility must be accounted for in a complete theory of function loss. But the theory of function loss described so far does not account for function loss without replacement. Function replacement is a special case of function loss. How can the account be improved to cover both function replacement and function loss without replacement?

Ideally, the condition for function loss without replacement should be a weakening of the condition on function loss with replacement. Function loss happens when some effect that was previously selected for ceases to be selected for. There are two ways that can happen. First, it can be that having the internal characteristics that enable that effect are being selected against. In this case there are some alternative internal characteristics that are being preferred by natural selection. These will amount to the cases of function replacement. Second, it can be that having the internal characteristics that enable that effect are being neither selected for nor selected against. This could happen, for example, when a reproducing trait's environment changes in such a way that having effect F (once a function) is no longer beneficial, but no alternative mutation has yet arisen that has an effect that can be selected for over the old standard (that is, all mutations so far do not improve success).

We have already mentioned vestigial traits; these are sometimes cases of function loss without replacement. Function loss without replacement can be accounted for by appeal to the fact that not all mutation types of a trait that are incapable of performing F are selected *against*. That is, there is function loss if a mutation type

arises that cannot have effect F and it is not selected against. If such a mutation type arises and it is selected for, we have function replacement. If such a mutation type arises and it is neither selected for nor selected against, then we have function loss without replacement.

The following objection helps make this claim more precise. Suppose an individual is born—say, an eagle—with a token trait that is unable to perform some function. Suppose its beak does not have the hooked tip that lets eagles tear their food up so easily. Let's furthermore suppose that this individual is also born with superior wings and eyesight, and these benefits outweigh the costs of its mutant beak, and so the eagle manages plenty of reproduction. This should not mean that the eagle beak has lost its food-tearing function. For it is the stronger wings and eyesight that are selected for. But the unhooked beak has failed to be selected against (it has been selected, but not selected for), so our condition seems to predict that there is function loss without replacement in this case.

In order to avoid this problem we need to distinguish selection that is not selection for from a failure to be selected against. The eagle's beak in our example is selected, but not selected for. But this beak is not an example of a failure to be selected against in the sense required by the theory of function loss. Whether an example of selection that is not selection for is an example of a failure to be selected against may only become apparent to us as scientists over several generations. We must watch and see what happens as the trait b (for beak) that was merely selected becomes separated from the traits e (for eyesight) and w (for wings) that were selected for. If b, when it occurs without e and w, tends to get selected against, then the beak of the individual who also had e and w was selected. But this instance of selection is not an instance of failing to be selected against in the sense required by the theory of function loss.

On the other hand, if b, when separated from e and w, does not tend to get selected against, then this instance of selection is an instance of failing to be selected against in the sense required by the theory of function loss, and so we would then say that the eagle's beak (amongst those eagles descended from individuals with b) has lost the food-tearing function.

With this distinction we can avoid the claim that the eagle's beak has lost the food-tearing function when a beak is unable to tear food but the individual with that beak has reproductive success as a result of other strengths.

This completes the theory of function loss. We see that homology is not a problematic criterion for individuating trait types.

3 The modal theory has insecure formal foundations

The modal theory of function is an alternative to the etiological theory that claims not to require a criterion for trait-type individuation, since it assigns function to

trait tokens, not trait types. The theory states that performing F is a function of an organism's token trait x at time t iff if x were F-ing at t, x's F-ing at t would contribute to the organism's inclusive fitness (Nanay 2010: 421f.). This section provides formal details that are absent in the literature surrounding the modal theory. According to Nanay (2010: 421), "[a]ny theory of counterfactuals could be used to fill in the details of [the modal theory], but, for simplicity" he opts for David Lewis's (1973) theory. In Sect. 3.1 we present Lewis' theory of counterfactuals and show how it is applied to spell out the modal theory of function; we begin with a nonformal discussion and introduce the formal apparatus in Sect. 3.1.1 and Sect. 3.1.2 (readers who prefer to neglect technical details may skip these sections). We then present two arguments to the effect that the modal theory has insecure formal foundations. In Sect. 3.2 we show that Lewis' centering assumption entails an inability to distinguish activities that yield benefits by accident from activities that are the performance of a function. We then consider two alternative formal theories that do not adopt the centering assumption and show that these also have untenable results when combined with the modal theory. In Sect. 3.3 we argue that under Lewis' theory of counterfactuals the modal theory either presupposes a criterion of trait-type individuation or is committed to a theory of transworld identity in cases of malfunction and dispositional functions in off-mode. If the commitment to transworld identity is problematic, this undermines the modal theory's principal motivation independently of the arguments presented in Sect. 2. Even if that commitment is not problematic, we see that adopting the modal theory yields commitments quite removed from the problems that theory was introduced to solve.

3.1 Lewis' theory of counterfactuals

Lewis' semantics employs possible worlds and similarity among possible worlds as basic notions in providing a semantics for counterfactuals. In order to check whether a counterfactual such as "If Bill had come to the party, it would have been fun" is true in the actual world, one goes to the possible world that is maximally similar to the actual world and where the antecedent is true (that is, the world as it would have been if Bill had come to the party) and checks whether or not the party was fun there. If it was, then the counterfactual is true; if it was not, then the counterfactual is false. In the following we use 'α' to abbreviate 'the actual world'. We will call a world where A is true 'A-world'.

Two complications are required though: first, it may not be that there is a single maximally similar world to α where the antecedent is true. Bill didn't come to the party, but if he had, would he have arrived at 19:45 or at 20:00? There may be a world w where Bill arrived at 19:45 and a distinct world w' where Bill arrived at 20:00, where both w and w' are equally good candidates for being the most similar

world to α where Bill came. That is, there may be ties for similarity. So we need to complicate the truth conditions for counterfactuals. In order to check whether a counterfactual $A\Box\!\!\rightarrow C$ (where A and C stand for sentences and $\Box\!\!\rightarrow$ stands for the semantic devices that conjoin those sentences into a counterfactual conditional) is true in α, one finds the *set* of possible world that are maximally similar to α and where the antecedent is true and checks whether the consequent is true at *all* those worlds. If it is, then the counterfactual is true; if it is not, then the counterfactual is false. We will call the members of the set of maximally similar antecedent worlds 'evaluation worlds' for any conditional with that antecedent.

Second, for some sentence A, there may not be a maximally similar world to α where A is true. This can happen in two ways: first, there may not be any A-worlds. For Lewis, any counterfactual with such an antecedent is vacuously true. Second, there may be an infinite sequence of A-worlds, each more similar to α than the last. In this case we cannot speak of the (set of) *maximally* similar world(s) to α. For example, Jim is 176 cm tall. Are there worlds maximally similar to α where Jim is over 180 cm tall? How tall is he in those worlds? If he is 180.5 cm in w, it seems that w', where he is 180.25 cm, is more similar to α (since Jim's height in w' is closer to his height in α than is his height in w). But then consider w'', where Jim is 180.125 cm. World w'' is more similar to α than is w'. We can continue this sequence infinitely, and never get to a *maximally* similar A-world.

This latter possibility complicates Lewis' truth conditions for counterfactuals substantially, but need not bother us for the purposes of this article. None of the antecedents we will consider are ones for which there is an infinite sequence of possible worlds, each more similar to α than the last, where the antecedent is true. So for each antecedent A we consider we may safely refer to the set of maximally similar A-worlds to α.

So a counterfactual $A\Box\!\!\rightarrow C$ is true just in case the maximally similar A-worlds are all C-worlds. On the modal theory, a trait-token x of an organism O has as function to F at time t iff if x were to F at t, then doing so would increase O's inclusive fitness. Hence a trait-token x of an organism O has as function to F iff at all maximally similar worlds where x is F-ing at time t this increases O's fitness. However, Nanay needs to add one additional condition. For example, on some occasion it might be true that if Bill were to fly, it would increase Bill's inclusive fitness. But we should not claim that flying is a function of Bill's on that occasion, at least not if Bill is a human. Nanay (2010: 422, 425) suggests that the closest possible worlds where Bill flies are too unlike α to be considered (and whether a possible world is worthy of consideration may vary according to a researcher's explanatory project). The modal theory must read:

The modal theory: A trait-token x of an organism O has as function

> to F iff there is a relatively similar world (given some explanatory project) where x is F-ing at time t and at all maximially similar worlds where x is F-ing at time t this increases O's fitness.

Next some comments about similarity. Lewis assumed that α is uniquely maximally similar to itself. That is, though ties for similarity are sometimes permitted, they are excluded in the case of α; no world is as similar to α as α is to itself. This is called the *centering* assumption. Centering ensures that when a counterfactual antecedent is true at α, the only world relevant for determining the truth of the counterfactual is α; the counterfactual is true if the consequent is true at α and false if the consequent is false at α.

The centering assumption is not unmotivated: even if there is a possible world w that is qualitatively indistinguishable but numerically distinct from α, still only α is numerically identical to α, and hence there is at least one respect in which α is more similar to α than w is. Nonetheless, Lewis considered weakening this assumption, allowing worlds that differ from α in some respect to be as similar to α as α is to itself. Still, no world is more similar to α than α is to itself. This is called the *weak centering* assumption. Under the weak centering assumption, when a counterfactual antecedent is true at α, α is one of the worlds relevant for determining the truth value of the counterfactual, but other worlds may be relevant as well.

Finally (though Lewis did not discuss this possibility), we might allow that for some antecedent A true at α, there are worlds that are *more* similar to α than α is to itself. We consider one such theory in Sect. 3.2.3. In this case, α need not be amongst the evaluation worlds for a counterfactual $A \mathrel{\Box\!\!\rightarrow} C$, even though A is true at α.

We now present Lewis' theory in formal detail; we resume nonformal discussion in Sect. 3.2.

3.1.1 Syntax

The language of Lewis' theory is given by the language of propositional logic with the addition of the $\Box\!\!\rightarrow$ operator for counterfactual implication. Let P be the set of atomic sentences. Then the complex sentences are defined inductively:

$$(1) \quad A \quad := \quad p \in P \;\Big|\; \neg A \;\Big|\; A \wedge C \;\Big|\; A \mathrel{\Box\!\!\rightarrow} C$$

In other words, sentences of Lewis' theory are either atomic sentences or complex sentences built from sentences using negation, conjunction, or counterfactual implication. We neglect disjunction and material implication here, leaving those to be defined in terms of negation and conjunction.

3.1.2 Semantics

We will simplify Lewis' theory: the semantics provided here will not in fact give the correct truth conditions for some kinds of embedded counterfactuals. These simplifications are adopted because they ease presentation of the view without harm to our purposes, as we are not concerned with any embedded counterfactuals here. Readers interested in a fully detailed semantics are referred to Lewis (1973). The sentences of Lewis' theory are interpreted in models $\mathscr{M} = \langle W, <, V \rangle$ where: (i) W is a non-empty set interpreted as the set of possible worlds; we stipulate that the actual world $\alpha \in W$. (ii) $<$ is a strict partial order on $W_\alpha \subseteq W$. W_α is interpreted as the set of possible worlds which are accessible from α. $<$ is interpreted as the comparative similarity relation with respect to α.[2] For example, for $w, w' \in W_\alpha$, $w < w'$ expresses that w is more similar to α than w'. (iii) For any atomic sentence $p \in P$, $V(p) \subseteq W$ is interpreted as the set of worlds $w \in W$ where p is true. The truth of a sentence at a world in a model of Lewis' theory can now be expressed as follows:

(2) $\mathscr{M}, w \vDash A \quad := \quad A$ is true at the possible world $w \in W$ in model $\mathscr{M}$

We will say that a world $w \in W$ in model $\mathscr{M}$ is a A-world in $\mathscr{M}$ if and only if $\mathscr{M}, w \vDash A$. The truth-conditions of negation and conjunction are standard; the truth-conditions of counterfactual implication are quite involved:

(3)
$\mathscr{M}, w \vDash p$ iff $w \in V(p)$
$\mathscr{M}, w \vDash \neg A$ iff not $\mathscr{M}, w \vDash A$
$\mathscr{M}, w \vDash A \wedge C$ iff $\mathscr{M}, w \vDash A$ and $\mathscr{M}, w \vDash C$
$\mathscr{M}, \alpha \vDash A \mathbin{\square\!\!\rightarrow} C$ iff $\forall w \in W_\alpha$ such that $\mathscr{M}, w \vDash A : \exists w' \in W_\alpha$ such that $\mathscr{M}, w' \vDash A$ and $w' \leq w$ and $\forall w'' \in W_\alpha$ such that $\mathscr{M}, w'' \vDash A$ and $w'' \leq w' : \mathscr{M}, w'' \vDash C$

The truth-conditions of counterfactual implication can be simplified by requiring that a model $\mathscr{M}$ is limited (Lewis 1973: 19f.):

(4) $\mathscr{M}$ is limited iff the strict partial order $<$ is well-founded

If a model $\mathscr{M}$ is limited, the set of accessible worlds $W_\alpha \subseteq W$ has at least one world $w \in W_\alpha$ which is maximally similar to α. In other words, infinite chains of ever more similar possible worlds in W_α are excluded. For any limited model $\mathscr{M}$, let $M \subseteq W_\alpha$ be the set of maximally similar possible worlds to α:

2 Herein lies our simplification. We have defined $<$ and accessibility only relative to α. A fully general semantics would define an accessibility relation and a comparative similarity relation for *every* possible world $w \in W$ in any model $\mathscr{M}$.

(5) $w \in M$ iff $\neg\exists w' \in W_\alpha$ such that $w' < w$

By the same token, if there is an A-world, then there is a nonempty set of maximally similar possible A-worlds $M_A \subseteq W_\alpha$ in $\mathscr{M}$:

(6) $w \in M_A$ iff $\mathscr{M}, w \vDash A$ and $\neg\exists w' \in W_\alpha$ such that $w' < w$

Assuming all models are limited, the truth-conditions of counterfactual implication can be simplified as follows (Lewis 1973: 20):

(7) $\mathscr{M}, \alpha \vDash A\Box\!\!\rightarrow C$ iff $\forall w \in M_A : \mathscr{M}, w \vDash C$

In other words, $\mathscr{M}, \alpha \vDash A\Box\!\!\rightarrow C$ iff all maximally similar A-worlds in $\mathscr{M}$ are also C-worlds. The modal theory of function states that performing F is a function of an organism O's token trait x at time t iff, if x were F-ing at t, x's F-ing at t would contribute to O's inclusive fitness, and there is a relatively similar world (given some explanatory project E) where x were F-ing at t. Let $R \subseteq W_\alpha$ be the set of relatively similar possible worlds in a model $\mathscr{M}$ given some explanatory project E:

(8) $w \in R$ iff $w \in W_\alpha$ and w is similar enough to α given E

The modal theory can now be succinctly stated as follows, where A abbreviates 'x is F-ing at t' and C abbreviates 'x's F-ing contributes to O's inclusive fitness':

(9) $\mathscr{M}, \alpha \vDash$ performing F is function of x at t iff (i) $\forall w \in M_A : \mathscr{M}, w \vDash C$
(ii) $M_A \cap R \neq \emptyset$

In other words, to F is a function of a token trait x at time t at the actual world α in a model $\mathscr{M}$ iff $\mathscr{M}, \alpha \vDash A\Box\!\!\rightarrow C$ and $M_A \cap R \neq \emptyset$ iff all A-worlds in $\mathscr{M}$ that are maximally similar and relatively similar to α are also C-worlds.

3.2 Are all useful behaviors functions?

The modal theory states that performing F is a function of an organism O's token trait x at time t iff if x were F-ing at t, x's F-ing at t would contribute to the organism's inclusive fitness (Nanay 2010: 421). But Nanay does not wish to claim that every instance of a token trait's activity that actually increases inclusive fitness is the performance of a function; sometimes behaviors increase fitness by accident:

> It would indeed be a worrying consequence of my view if it ended up assimilating function to use [...]. But this is not the case. What a trait is being used for is determined by *what goes on in the actual world*. Function (and, arguably, usefulness), in contrast, depends on *what goes on in nearby possible worlds*. Function is a modal concept; use is not (Nanay 2010: 427, emphasis added).

In Sect. 3.2.1 we show that adopting Lewis' semantics yields this undesired result. In Sect. 3.2.2 and Sect. 3.2.3 we consider alternatives to Lewis' semantics and show that the problem cannot be fully resolved by adopting either of these options. This does not exhaust the space of logical possibilities, but it does exhaust the options that are obvious to us.

3.2.1 Lewis' centering assumption

Models in Lewis' theory of counterfactuals are centered (see Lewis 1973: 14f.). This means that the actual world is uniquely maximally similar to itself. This has as consequence that when a counterfactual's antecedent is true at the actual world, the actual world is the only evaluation world. So the truth value of the counterfactual is equal to the actual truth value of the consequent. In other words, if a counterfactual's antecedent is true, the counterfactual reduces to a material conditional.

Now integrate this with the modal theory. Suppose that some trait token x is having effect F at time t. Given Lewis' semantics, the modal theory will count this F-ing as a function if and only if x's F-ing at t contributes to O's inclusive fitness (assuming that the actual world is always a relatively close world; we will persist in this assumption). So we will not be able to distinguish accidents that increase inclusive fitness from functions.

Return to our example of a behavior that accidentally leads to increase in inclusive fitness on some occasion. Suppose Roger the rabbit is scared by a falling branch and thumps his leg, thereby scaring conspecifics. Suppose there happens to be an undetected predator nearby that would have eaten a conspecific if Roger had not thumped. Thus Roger's thumping increases his inclusive fitness on this occasion.

In the actual world, Roger's thumping contributes to inclusive fitness. On Nanay's modification of Lewis' semantics with centering, the counterfactual, 'If Roger had thumped on that occasion, doing so would have contributed to his inclusive fitness' is true. Given the modal theory of function, then, Roger had as function to thump on that occasion. But Roger's thump was a paradigm example of an increase in inclusive fitness by accident, not by function performance. So on this semantics the modal theory cannot account for the difference between function performance and behaviours that accidentally increase inclusive fitness. Using Nanay's expression,

the theory assimilates function to use.

Now we state the argument formally. For any model $\mathscr{M}$:

(10) $\mathscr{M}$ is centered iff $\forall w \in W_\alpha$ such that $w \neq \alpha, \alpha < w$

Consider the following argument-schema (call it 'assimilation-schema') with a behavior F, an organism O's token trait x, a time t and any centered model $\mathscr{M}$:

P1 $\mathscr{M}, \alpha \vDash A$
P2 $\{\alpha\} \cap R \neq \emptyset$
C1 By P1, P2 and centering, $M_A = \{\alpha\}$ and hence $M_A \cap R = \{\alpha\}$
C2 By C1 and the truth-conditions for counterfactuals, $\mathscr{M}, \alpha \vDash A \mathbin{\square\!\!\rightarrow} C$ only if $\mathscr{M}, \alpha \vDash C$
C3 By C2 and the modal theory of function, $\mathscr{M}, \alpha \vDash$ performing F is a function of x at t only if $\mathscr{M}, \alpha \vDash x$'s F-ing contributes to O's inclusive fitness

From this argument we conclude that the modal theory supplemented by Lewis' semantics is unable to distinguish functions from actions that generate increased inclusive fitness by accident. The next two sections explore two methods of rejecting the centering requirement that have been offered in the literature, and show that neither yields an adequate semantics that does the work Nanay requires.

3.2.2 Weak Centering

This section explores whether the modal theory can distinguish between functions and accidents that increase inclusive fitness by employing a *weakly* centered semantics for counterfactuals (see Lewis 1973: 29). On this assumption, there may be worlds that are distinct from the actual world α but that are as similar to α as α is to itself (call such worlds *center worlds*). This opens the possibility that a counterfactual $A \mathbin{\square\!\!\rightarrow} C$ is false in α even though A and C are both true in α.

The modal theory with a centered semantics cannot distinguish accidents that increase inclusive fitness from functions that are actually performed because if the performance of an action increases inclusive fitness at α, then the counterfactual 'If that action had been performed, it would have increased inclusive fitness' is true and so that action is a function according to the modal theory. Take again the example of the rabbit's leg thump and consider two cases. In the first, at the actual world, the rabbit thumps its leg due to the detection of a nearby predator. In the second, at the actual world, the rabbit is thumps due to a falling branch and there is an undetected predator nearby. In both cases the thumping increases inclusive fitness. However, the modal theory should attribute the function of warning conspecifics to the thumping in the first case but not in the second. In the latter inclusive fitness is increased only

by accident. On a centered semantics, this requirement is not met, as we have shown in Sect. 3.2.1.

On a weakly centered semantics, the function ascription in the accidental second case can be blocked only by admitting a center world where the rabbit thumps due to a falling branch but there is no undetected predator that would otherwis have eaten a conspecific nearby (call such a world 'blocking world'). This renders false the counterfactual 'If the rabbit had thumped, it would have increased inclusive fitness': the evaluation worlds comprise the center worlds and the consequent is not true at all center worlds. Hence, the function of warning conspecifics is not ascribed to the rabbit's leg-thumping, as wanted.

For this solution to work, there must be a noncircular means of specifying the similarity relation that guarantees the existence of a blocking world in cases of accidental increases in inclusive fitness and that guarantees the nonexistence of a blocking world in cases of function performance. We offer a proposal: let the center worlds be those where all nonaccidental generalizations of the actual world hold.[3] At first glance this might seem plausible: the very accidentality of accidental increases in inclusive fitness would be accounted for by the existence of a center world where the increase in inclusive fitness does not obtain. Our skittish rabbit increased fitness accidentally by thumping in response to a falling branch when there was an undetected predator nearby, but that the branch fell when there was a predator nearby was an accident, and so there will be a center world where the branch does not fall when the predator is nearby: in some, because the branch does not fall; in others because the predator is not there. Presumably in the worlds where the branch doesn't fall, the rabbit does not thump; these worlds are not amongst the evaluation worlds. But the worlds where the branch does fall but the predator is not there are relevant; in those worlds there is no increase in inclusive fitness and so the counterfactual 'if the rabbit had thumped, it would have increased inclusive fitness' is false at α.

But on closer inspection this principle is too strong. It (plausibly) makes all the counterfactuals associated with accidental increases in inclusive fitness (teh *accidental* counterfactuals) come out false by making it very easy for counterfactuals with true antecedents to come out false. Too easy: the counterfactuals associated with instances of function performance (the *function* counterfactuals) will too often come out false. Take a case where a rabbit observes a predator, and so thumps, warns conspecifics, and increases inclusive fitness. Here it is not accidental that the rabbit thumped when the predator was there. Still, it might be accidental that the predator was there at all. That means that there will be center worlds where the predator is not there. And at some of those worlds, some other accident causes the rabbit to thump

3 Not all authors of this paper are convinced that this distinction can be sensibly drawn, but all are willing to admit the possibility for the sake of argument.

(say, a falling branch). Then the function counterfactual, 'If the rabbit had thumped, it would have increased inclusive fitness' is false, since there is a center world where the rabbit thumps, there is no predator around, and so there is no resulting increase in inclusive fitness.

We will not discuss alternative principles that might generate the similarity orderings needed by the modal theory to distinguish actual function performances from accidental increases in inclusive fitness. We take this argument to sufficiently demonstrate the challenges involved: the principle must make it easy enough for a counterfactual to come out false that all the accidental counterfactuals come out false; and must make it difficult enough for a counterfactual to come out false that all the function counterfactuals come out true. We see no principle that can strike this balance.

We now state this argument formally. For any model $\mathscr{M}$:

(11) $\mathscr{M}$ is weakly centered iff $\neg\exists w(w \in W_\alpha \text{ and } w < \alpha)$

A semantics that replaces centering with weak centering undermines the objection of Sect. 3.2.1 by invalidating the inference to C1 in the assimilation schema. However, it gives rise to a new schema (call it 'blocking-schema') with a behavior F, an organism O's token trait x, a time t and any centered model $\mathscr{M}$. We use '$\simeq$' to indicate the relationship that holds between w and w' iff w is as similar to w' as w' is to itself:

P1 $\mathscr{M}, \alpha \vDash A \wedge C$
P2 $\{\alpha\} \cap R \neq \emptyset$
C1 By P1 and weak centering, $M_A = \{w \in W_\alpha : w \simeq \alpha\}$ and hence $M_A \cap R = \{w \in W_\alpha : w \simeq \alpha\}$[4]
C2 By C1 and the truth-conditions for counterfactuals, $\mathscr{M}, \alpha \vDash A \mathbin{\square\!\!\rightarrow} C$ only if $\forall w \in M_A : \mathscr{M}, w \vDash C$
C3 By C2 and the modal theory of function, $\mathscr{M}, \alpha \vDash$ performing F is a function of x at t only if $\forall w \in M_A : \mathscr{M}, w \vDash x$'s F-ing contributes to O's inclusive fitness

Accidental cases of function ascription can only be prevented by admitting a blocking world which renders false the necessary condition specified in C3:

(12) $w \in M_A$ is a blocking world iff $\mathscr{M}, w \nvDash x$'s F-ing contributes to O's inclusive fitness

4 We assume that any world that is as similar to α as α itself is a relatively similar world according to any explanatory project.

But we see no principle that will guarantee that a blocking world exists when required by the modal theory and does not exist when the modal theory requires there not be one. So we see no satisfactory solution to our problem by adopting a weakly centered semantics for counterfactuals.

3.2.3 Iatridou-style semantics

An second way to weaken the centering requirement is proposed in Iatridou (2000), which we will call the *exclusion* requirement. On this view (again neglecting embedded conditionals) the actual world is always excluded from the set of maximally similar antecedent worlds. That is to say, a counterfactual is evaluated by checking whether the consequent is true at all maximally similar worlds to α and where the antecedent is true, except α. So if α is a world where the antecedent is true, it is by stipulation excluded from the evaluation worlds.

The problem with this semantics is that it either invalidates modus ponens[5] or encounters one of the problems described in Sect. 3.2.1 and Sect. 3.2.2. We argue by considering three cases that are jointly exhaustive. In the first case, every evaluation world differs from the actual world in the truth of some sentence. The second case drops this assumption but requires that the worlds that make exactly the same sentences true as the actual world are more similar to the actual world than is any world that differs from the actual world on the truth of some sentence. The third case drops both assumptions. In the first case modus ponens is invalidated; in the second case the problem of Sect. 3.2.1 is encountered; in the third case the problem of Sect. 3.2.2 is encountered.

Case 1. Consider some sentence A $\in P$ that is true at α. We construct a counterexample to modus ponens as follows. Suppose that the most similar worlds to α where A is true but that exclude α are the members of the set $M_A \setminus \{\alpha\} = \{w_1, w_2, \ldots, w_n\}$. Each of these possible worlds differs from the actual world in the truth of some sentence. Thus we have a set of sentences: the set of sentences $X \in P$ such that for some world $w_x \in M_A \setminus \{\alpha\}$, X is true at w_x and false at α. For example, sentence 1 is the sentence that is true at w_1 and false at α; sentence 2 is the sentence that is true at w_2 and false at α; and so on. Now consider the disjunction of these sentences, $(1 \vee 2 \vee \ldots \vee n)$. This sentence is true at every evaluation world, but false at α. As a result, the counterfactual A$\Box\!\!\rightarrow(1 \vee 2 \vee \ldots \vee n)$ is true at α, as is A; yet $(1 \vee 2 \vee \ldots \vee n)$ is not true at α. Thus modus ponens is invalidated.

Case 2. Now consider the case where we drop the assumption that every eval-

5 McGee (1985) and Lycan (2001) argue that modus ponens is invalid anyway, though perhaps weakenings of modus ponens hold, such as modus ponens restricted to conditional premises that do not nest other conditionals. Iatridou's semantics also invalidates this weaker principle, though we will not discuss those principles here.

uation world differs from the actual world with respect to the truth value of some sentence. If A is true at α and there are worlds that do not differ from α in the truth of any sentence, then if the worlds that do not differ from α in any respect are the most similar worlds to α where A is true (excluding α), we face the problem of assimilating function to use as described in Sect. 3.2.1. For although the actual world is not an evaluation world, the evaluation worlds make all the same sentences true as does the actual world, and so the counterfactual A$\Box\!\!\rightarrow$C will be true if and only if C is true at α.

Case 3. On the other hand, it might be that there are worlds that differ from the actual world in the truth of some sentence but that are no less similar to the actual world than the worlds that do not differ from the actual in the truth of any sentence. Here we face again the problem addressed in Sect. 3.2.2. We do not invalidate modus ponens, and we do not face the problem of assimilating function to use head-on, but we do face the challenge of providing a noncircular principle that establishes that there is a blocking world in cases of accidental increases in inclusive fitness and that there is not a blocking world in cases of function performance.

We conclude that the problem of assimilating function to use cannot be solved by employing Iatridou-style semantics for counterfactuals.

The arguments of this section—the rejection of a weakly centered semantics and Iatridou-style semantics as means to distinguish between functions and accidents that increase inclusive fitness—have not been exhaustive. But we can note a challenge that Nanay must face. If Nanay wishes to avoid the assimilation of function to use while maintaining the modal theory, he must offer a semantics on which the counterfactual does not entail the material conditional. For in cases of actions that increase inclusive fitness by accident, the material conditional 'If that action was performed then the performance to that action yielded increased inclusive fitness' is true. But Nanay must maintain that the counterfactual 'If that action had been performed, the performance of that action would have yielded an increase in inclusive fitness' is false. While few modern theorists maintain that the semantics of the counterfactual is the semantics of the material conditional, few deny that the counterfactual entails the corresponding material conditional. Thus the space of options available to the modal theory is seriously constrained.

3.3 Transworld identity or trait-type dependence

The principal virtue of the modal theory is its trait-type independence. However, there are two classes of cases that show that the modal theory either presupposes a criterion of trait-type individuation or is committed to transworld identity. First are cases of malfunction; second are cases of dispositional functions in off-mode.

A theory of function must account for malfunction (Nanay 2010: 414). If x

malfunctions with respect to F at t at the actual world, then x fails to F at α. For example, if Obama's heart malfunctions, then it fails to pump blood. A theory of function must also account for dispositional functions in off-mode. Many functions are not performed constantly but only under specific circumstances. For example, the tongue of the juvenile aquatic garter snake *Thamnophis atratus* has the function to lure prey but it lures prey only when the snake is hunting (Welsh & Lind 2000). The p53 gene has the function to brake the cycle of cell growth but it brakes the cycle of cell growth only in circumstances where the cell is stressed or damaged (Bert Vogelstein & Levine 2000). Dispositional functions have at least two modes (on and off). If x has a dispositional function F in off-mode at t in α, then x is not F-ing at α.

The modal theory aims to account for malfunction as follows. For a function F, an organism O's token trait x and a time t, x malfunctions at t iff x has function F and x fails to F. That is, if x were to F at t, doing F would increase O's inclusive fitness; but x in fact fails to F (Nanay 2010: 422).

To see that the modal theory either is committed to the notion of transworld identity or presupposes a criterion of trait-type individuation, consider a case where an organism O's trait x has the function to perform F at t at α but x malfunctions. Then x does not F at α but there exists a set $\{w_1, w_2, ...w_n\}$ of maximally similar relatively similar possible worlds where x Fs at t and x's F-ing at t contributes to O's inclusive fitness.

In this case, the function F of the token trait x at α depends on what goes on in $w_1, w_2, ...w_n$. Let us call x at α 'x_α', x at w_1 'x_{w_1}', x at w_2 'x_{w_2}', and so on. Now there are two jointly exhaustive cases with respect to the identity of x_α, $x_{w_1}, ..., x_{w_n}$.

Case 1. Some of x_α, $x_{w_1}...x_{w_n}$ are nonidentical; nonidentical individuals from this group may be called *counterparts*. The problem of trait type individuation arises since the function of x_α is determined by the properties of $x_{w_1}...x_{w_n}$. For example, if Obama's heart is malfunctioning at α, then the function of his heart is determined by whether the closest relatively close hearts (and not: lungs, brains or ears) are pumping blood and contributing to inclusive fitness. So a criterion of identifying hearts across worlds is required. This must enable us to distinguish the counterparts of Obama's heart from the counterparts of his lungs, brain, ears, and so on. That is, we need a criterion for individuating trait types across possible worlds.

Case 2. All of x_α, $x_{w_1}...x_{w_n}$ are identical. In other words, x_α is part of an organism O living at α and $w_1...w_n$ and $x_\alpha, x_{w_1}...x_{w_n}$ are parts of the same organism O living at all the same worlds. This is transworld identity. Here the modal theory avoids the problem of trait-type individuation: the function of x_α is determined entirely by the properties of x_α since $w_1, w_2, ...w_n$ are the most similar relatively similar possible worlds where x is F-ing, and x_α, x_{w_1},...x_{w_n} are identical. We remain agnostic regarding whether problems for the theory of transworld identity are solvable. We

want to underscore that the modal theory is forced to take a stance in an intricate debate that is distinct from the original problem of analyzing biological function.

In sum, for cases of malfunction, the modal theory must either adopt the theory of transworld identity or lose its principal selling point. The problem for dispositional functions in off-mode is parallel. Suppose a juvenile aquatic garter snake is not in fact hunting; then whether its tongue has the function of luring prey is determined by either (a) what counterpart tongues do in nearby possible worlds (hence the problem of trait type individuation) or by (b) what its self-identical tongue is doing in other possible worlds (hence a commitment to transworld identity). It might be objected that there are no strictly dispositional functions, only functions at times; this would undermine our objection with respect to dispositional functions. But the problem stands with respect to malfunction.

4 Conclusion

We have offered a theory of function loss and shown how this theory undermines the problem of trait type independence for the etiological theory of function. Then we demonstrated that the modal theory faces internal challenges. First, it is not clear that there is a semantics for counterfactuals that enables the modal theory to distinguish function performance from accidental increases in inclusive fitness. Second, the modal theory cannot account for either malfunction or dispositional functions in off-mode while both avoiding commitment to transworld identity and maintaining trait type independence.

References

Bert Vogelstein, David Lane & Arnold J. Levine. 2000. Surfing the p53 network. *Nature* 408(6810). 307–310.

Gatesy, Stephen M. & Kenneth P. Dial. 1996. Locomotor modules and the evolution of avian flight. *Evolution* 50(1). 331–340.

Iatridou, Sabine. 2000. The grammatical ingredients of counterfactuality. *Linguistic Inquiry* 31. 231–270.

Kiritani, Osamu. 2011. Function and modality. *The Journal of Mind and Behavior* 32(1). 1–4.

Lewis, David. 1973. *Counterfactuals*. Malden: Blackwell.

Lycan, William. 2001. *Real conditionals*. New York: Oxford.

McGee, Vann. 1985. A counterexample to modus ponens. *Journal of Philosophy* 82.

Millikan, Ruth. 1984. *Language, thought, and other biological categories*. Cambridge: MIT.

Nanay, Bence. 2010. A modal theory of function. *Journal of Philosophy* 107(8). 412–431.

Nanay, Bence. 2011. Function, modality, mental content. *The Journal of Mind and Behavior* 32(1). 84–87.

Nanay, Bence. 2012. Function attributions depend on the explanatory context. *Journal of Philosophy* 109(10). 623–627.

Neander, Karen & Alexander Rosenberg. 2012. Solving the circularity problem for functions. *The Journal of Philosophy* 109(10). 613–622.

Welsh, Hartwell H. & Amy J. Lind. 2000. Evidence of lingual-luring by an aquatic snake. *Journal of Herpetology* 34. 67–74.