Erkenntnis (2021) 86:1045–1070
https://doi.org/10.1007/s10670-019-00143-0

ORIGINAL RESEARCH

The Explanatory Role of Concepts

Samuel D. Taylor[1] · **Gottfried Vosgerau**[1]

Received: 6 August 2018 / Accepted: 17 June 2019 / Published online: 2 July 2019

Abstract
Machery (Doing without concepts, Oxford University Press, New York, 2009) and Weiskopf (Synthese 169:145–173, 2009) argue that the kind CONCEPT is a natural kind if and only if it plays an explanatory role in cognitive scientific explanations. In this paper, we argue against this explanationist approach to determining the natural kind-hood of CONCEPT. We first demonstrate that hybrid, pluralist, and eliminativist theories of concepts afford the kind CONCEPT different explanatory roles. Then, we argue that we cannot decide between hybrid, pluralist, and eliminativist theories of concepts, because each endorses a different, but equally viable, specification of the explananda of cognitive science. It follows that an explanationist approach to determining the natural kind-hood of CONCEPT fails, because there is no consensus about whether or not CONCEPT should be afforded an explanatory role in our best cognitive scientific explanations. We conclude by considering what our critique of explanationism could imply for further discussions about the explanatory role of concepts in cognitive science.

1 The Explanatory Challenge to CONCEPT

According to the "received view," all concepts have a number of properties in common: they all store a single kind of information, they all have the same functional properties, and they are all acquired by the same type of learning process, etc. (Machery and Seppälä 2011, p. 99). On this view, a theory of concepts aims to describe these properties and so to account for the formation and application of concepts. Moreover, from the perspective of the received view, concepts are of one kind—the kind CONCEPT—and they "explain the properties of our higher cognitive competences"; that is, the properties of higher cognition that are operative in

✉ Samuel D. Taylor
sam.taylor@hhu.de; samuel.da.taylor@gmail.com

Gottfried Vosgerau
vosgerau@hhu.de

[1] Department of Philosophy, Heinrich-Heine-University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

cognitive tasks such as categorisation, meaning extraction, and inductive and deductive reasoning (Machery and Seppälä 2011, p. 99).[1]

In recent years, however, psychologists have identified several distinct types of categorisation judgements, several distinct operations of meaning extraction, and several distinct episodes of inductive/deductive reasoning, and so have been forced to posit a number of different representational kinds, with incommensurable properties, to do the required explanatory work. For example, PROTOTYPES (Rosch 1973; Lakoff 1987), BUNDLES OF EXEMPLARS (Nosofsky 1988), THEORY-LIKE STRUCTURES (Carey 1985; Gopnik and Meltzoff 1997; Rehder 2003), PERCEPTUAL 'PROXYTYPES' (Prinz 2002), and UNSTRUCTURED ATOMIC SYMBOLS (Fodor 1994) have all been posited to explain, say, categorisation and reasoning. As a result, it has been argued that the received view of concepts cannot be correct, because the explanatory work is not done by the single kind CONCEPT, but by a set of representational kinds that do not store the same kinds of information or have the same functional properties (Bloch-Mullins 2018).

To make the explanatory challenge to the received view concrete, consider explanations of categorisation judgements. In some cases, psychologists explain the categorisation of an individual *c* in a category *C* in terms of a correspondence between the properties of *c* and the typical properties of members of *C*. In cases such as these, psychologists posit the kind PROTOTYPE to do the required explanatory work. However, in other cases psychologists explain the categorisation of an individual *c* in a category *C* in terms of a judgement that *c* is sufficiently similar to salient members of *C*. In cases such as these, psychologists posit the kind EXEMPLAR to do the required explanatory work. According to the explanatory challenge, because psychologists are required to posit different representational kinds with different properties and functions to explain these different types of categorisation judgements, the kind CONCEPT is redundant in explanations of categorisation. And because the received view of CONCEPT fails to predict this explanatory diversity it is argued that it cannot be correct (cf. Machery and Seppälä 2011, p. 99).[2]

In this paper, we argue against one possible interpretation of the explanatory challenge: that the explanatory challenge demonstrates that CONCEPT is not a natural kind. This involves rejecting what we will call an explanationist approach for determining the natural kind-hood of CONCEPT, whereby CONCEPT is taken to be a natural kind iff CONCEPT features in a proposition best explaining some explananda of

[1] In this paper, we denote kinds with small caps.

[2] One quick and easy way to respond to the explanatory challenge to the received view is to argue that *all* posited representational kinds are of the super-kind CONCEPT in virtue of the fact that they are all constituents of thought. According to this response, all of the representational kinds posited to do explanatory work in cognitive science fall under the concept of concept in virtue of being conceptual in kind (Weiskopf 2009, pp. 147–148). Many philosophers have been quick to accept this view as part of their response to the explanatory challenge. However, even if one thinks that this response is broadly correct, this still leaves open a further question: what are the defining properties of the super-kind CONCEPT that make it the case that all representations doing explanatory work in cognitive science can be classified as concepts? Providing an answer to this question has proven to be controversial and remains the central issue of contention between competing theories of concepts that we discuss in sections three, four, and five below.

cognitive science. For example, according to an explanationist approach for determining the natural kind-hood of CONCEPT, if CONCEPT features in a proposition best explaining some aspects of categorisation—e.g., 'categorisation judgements involve the comparison of some individual *c* with some CONCEPT *C*'—, then CONCEPT is a natural kind.

Our argument against this explanationist approach to determining the natural kind-hood of CONCEPT runs as follows. In Sect. 2, we introduce the explanationist approach to determining whether or not CONCEPT is a natural kind. In Sects. 3 and 4, we introduce eliminativist, pluralist, and hybrid theories of concepts, and demonstrate that they each endorse different views of the explanatory role CONCEPT in cognitive science. In Sect. 5, we argue that the reason that eliminativist, pluralist, and hybrid theories of concepts endorse different views of the explanatory role CONCEPT is because they endorse different specifications of the explananda of cognitive science. We defend this claim from possible objections in Sect. 6. Then, in Sect. 7, we argue that because the explananda of cognitive science cannot be pre-theoretically specified, we cannot decide between eliminativist, pluralist, and hybrid interpretations of the explanatory role of CONCEPT. It follows that an explanationist approach for determining the natural kind-hood of CONCEPT cannot be made to work. We conclude by considering what our critique of explanationism could imply for further discussions about the explanatory role of concepts in cognitive science.

2 Explanationism About CONCEPT

The debate about how best to determine the natural kind-hood of theoretical posits is long and convoluted.[3] However, one approach that has found favour in recent times has been an explanationist approach to natural kind-hood determination. Explanationism holds that "a person's evidence supports a proposition just in case that proposition is part of the best available explanation for the person's evidence" (Byerly 2013). In this way, explanationism dictates that one must assent to the truth of a proposition—say, 'water is H_2O' or 'electrons have an intrinsic angular momentum (spin) of $\frac{1}{2}$'—iff that proposition is part of the best available explanation for some relevant evidence—say, that water has a boiling point of 100° or that paramagnetic substances (e.g., aluminium and oxygen) are weakly attracted to an applied magnetic field. Taken at the level of the truth-conditions alone, explanationism concerns only the validity of our beliefs and so appears to have nothing to say about the natural

[3] Much of the debate about theoretical posits has focused on the issue of our epistemic commitment to the putative referents of theoretical terms. A number of formal representations of the semantics of theoretical terms have been developed, which aim to make explicit the role of such terms in scientific theories. Perhaps the most famous of these formal frameworks are Ramsey sentences, which introduce a division between the set V_o of observational terms and the set V_t of theoretical terms in a theory (Ketland 2004; Ramsey 1931). The framework of Ramsey sentences trades on the idea that theories are syntactic structures that can be parsed in terms of a language $L(V_o, V_t)$. However, it is also possible to formally represent the role of theoretical terms on a semantic conception of theories, albeit then in terms of, e.g., partial structures or modal relations (cf. Suppes 1967, 2002; Van Fraassen 1980).

kind-hood of theoretical posits. However, the explanationist line of argument can be taken further in light of the notion that our best explanations are "the only workable criterion of reality" (Ellis 2001).

Explanationism can be seen as imposing a condition on the reality of any entity posited in scientific explanation, because our commitment to any posit x can be assessed in accordance with whether or not x features in a proposition that best explains some piece of evidence (Saatsi 2017). When tied to a broadly scientistic picture, this kind of explanationism proposes that we should be committed to all posits x that feature in propositions that best explain the evidence to be accounted for by science. So, on this view, we should be committed to the posits H_2O and *Electron*, because H_2O and *Electron* feature in at least one proposition that best explains some subset of the evidence to be accounted for by science. However, we should not be committed to the posit *Witches*, because *Witches* does not feature in any proposition that best explains any evidence to be accounted for by science. Thus, this application of explanationism holds that "something is real if its positing plays an indispensable role in the explanation of well-founded phenomena" (Psillos 2005, p. 398). In a slogan: some posit x exists iff taking x to exist helps us to formulate scientific explanations of something that would otherwise be puzzling (Suppes 2008, p. 16).[4]

An explanationist approach to natural kind-hood determination purports to identify those posits that are the best candidates for the kinds that carve reality at its joints. So, by applying the explanationist approach to natural kind-hood determination, we expect to arrive at some principled demarcation of the kinds that are 'natural'—say, the kinds GOLD and ELECTRON—and the kinds that are not—say, PHLOGISTON and GRIFFIN (Kitcher 1993).[5] The difficulty, however, is that in order to successfully apply the explanationist approach to natural kind-hood determination, we first have to specify the evidence that is to be explained. The question of how to specify the relevant 'evidential explananda' is the root cause of much philosophical wrangling about, for instance, the reduction of the special sciences to the fundamental sciences or the reality of abstract and/or fictional objects. But putting aside these large and cumbersome questions, one thing can be made clear: if we can agree that there are some evidential explananda to be explained by science, then it follows that an explanationist approach to natural kind-hood determination ought be able to sort out the

[4] One may argue that an explanationist approach to natural kind-hood determination is inadequate, because best explanations often invoke categories that are not to be thought of as natural kinds, even where these categories are useful, causally significant, partially explanatory, etc. This concern, however, is with the explanationist idea that natural kinds can be somehow read off of our best explanations. Although this concern may well be valid, our approach in this paper puts this potential problem with explanationism to one side. We accept—for the purposes of our argument—the explanationist claim that all posits in our best explanations are natural kinds. However, we question the viability of this explanationist methodology with respect to determining the natural kind-hood of CONCEPT. Thus, as we will demonstrate below, our argument does not turn on the question of whether or not CONCEPT can be thought of as natural kind when posited in our best explanations, but, rather, whether or not we can be sure that CONCEPT will be posited in our best explanations at all.

[5] Consider here Bird and Tobin's (2017) general definition of a natural kind as a kind that "corresponds to a grouping that reflects the structure of the natural world rather than the interests and actions of human beings".

natural from non-natural kinds according to which kinds are posited in explanatory propositions that best explain the evidential explananda.[6]

Our concern in this paper is with the evidential explananda of cognitive science. We assume, therefore, that there is something for cognitive science to explain. In short, we hold that the evidential explananda of cognitive science is any evidential explananda that reveals the functioning and operation of cognition or the achievement or undertaking of cognitive competencies more generally. That is, we assume that the evidential explananda of cognitive science are any evidential explananda that concern the operations of the mind. With this proviso out of the way, we can apply the explanationist approach to natural kind-hood determination to the kinds commonly associated with cognitive science. For example, we can apply the explanationist approach to natural kind-hood determination to kinds such as REPRESENTATION, NEURON, or MODULE. What's more, we can apply the explanationist approach to natural kind-hood determination to perhaps the most prominent kind associated with cognitive science: the kind CONCEPT. An explanationist approach to determining the natural kind-hood of CONCEPT holds that the kind CONCEPT is a natural kind if and only if there is a proposition featuring CONCEPT that is part of the best available explanation for some relevant evidence to be explained by cognitive science. For example, if the proposition, 'CONCEPT's are constituents of thought,' is part of the best explanation of some relevant evidence to be explained by cognitive science, then we would be justified in supposing that CONCEPT is a natural kind. In this way, the explanationist approach to determining the natural kind-hood of CONCEPT indexes the natural kind-hood of CONCEPT to CONCEPT's explanatory role in cognitive science.

In recent debates about concepts, an explanationist approach to determining natural kind-hood of CONCEPT has come to fore. For example, both Machery (2009) and Weiskopf (2009) endorse an explanationist approach to determining the natural kind-hood of CONCEPT, even if they have not phrased their views in exactly these terms. Weiskopf (2009, p. 147) argues that the kind CONCEPT—like all other kinds that are worthy of our interest—is to be understood as a "groupings of entities that participate in a body of empirically discovered reliable generalizations, and which participate in those generalizations due to some set of properties they have in common." Similarly, Machery (2009, p. 232) argues that a class *C* of entities—for instance, the class *C* that constitutes the kind CONCEPT—"is a natural kind if and only if there is a large set of scientifically relevant properties such that *C* is the maximal class whose members tend to share these properties because of some causal mechanism"; where "scientifically relevant" is to be read as saying that the *C* is a class about which many explanatory generalizations can be formulated (cf. Machery 2009, p. 232, for further discussion of Machery's "Scientific Eliminativism" about natural kinds).[7] In

[6] We assume that the explananda of any given science are the phenomena that belong to the subject matter of that science. We accept, however, that the question of which science a phenomenon belongs to seems to be interconnected with the possibility that a given science has to explain the phenomenon in question. We discuss these points in greater detail in Sect. 7 below.

[7] It is relevant here to note that Machery endorses a specific characterisation of natural kinds as borrowed from, e.g., (Boyd 1991 ,1999). On this "causal notion of natural kind," a natural kind is "a class about which many generalizations can be formulated," because "its members tend to have many properties in common" and "there is at least one causal mechanism that explains why its members tend to

the remainder of this paper, we want to build the case against such an explanationist approach to determining natural kind-hood of CONCEPT.

3 Theories of CONCEPT

In the current literature about concepts, there is an ongoing discussion about how to conceive of the explanatory role of the kind CONCEPT given the explanatory challenge to the received view. In this section, we will use the notation of set theory to elucidate and compare the different theories of CONCEPT's explanatory role. Take *concept eliminativism* to start with (Machery 2009, 2010). For the eliminativist, there are potentially many different representations of one and the same thing. For example, for cats there could potentially be a cat-prototype representation, a cat-exemplar representation, and a cat-theory-like-structure representation. Thus, there could be at least three different tokens of mental representations standing for cats; namely p_{CAT} (the cat-prototype), e_{CAT} (the cat-exemplar), and t_{CAT} (the cat-theory-like-structure). The kind PROTOTYPE, for instance, can then be described as the set of all prototype tokens, where the defining property of this set is the (complex) property P of being a prototype, whatever that might be in detail:

$$PROTOTYPE = \{x \mid P(x)\} = \{p_{CAT}, p_{DOG}, \ldots\}$$

The eliminativist interpretation of the kind CONCEPT can then be construed in one of two ways: either as the set of all representational kinds with the defining property of figuring in scientific explanations of higher order cognitive capacities; or as the set of all representations with the complex property C_E, where C_E is no more than the exclusive disjunction of the different defining properties of the explanatorily valuable kinds.

$$\begin{aligned} CONCEPT_{E_1} =& \{x \mid x \text{ figuring in scientific explanations}\} = \\ =& \{\{x_1 \mid P(x_1)\}, \{x_2 \mid E(x_2)\}, \{x_3 \mid T(x_3)\}\} \end{aligned}$$

or

$$\begin{aligned} CONCEPT_{E_2} =& \{x \mid C_E(x), \text{ where } \forall y\big(C_E(y) \leftrightarrow (P(y) \vee E(y) \vee T(y))\big)\} = \\ =& \{p_{CAT}, e_{CAT}, t_{CAT}, p_{DOG}, e_{DOG}, t_{DOG}, \ldots\} \end{aligned}$$

The first set $CONCEPT_{E_1}$ is simply the set that contains all the different representational kinds as sets, which are in our example the set of all prototypes, the set of all exemplars, and the set of all theory-like structures. The second set $CONCEPT_{E_2}$ is the set that contains all tokens of the different representational kinds; so, it contains all prototypes (but not the set of all prototypes), all exemplars (but not the

Footnote 7 (continued)
have those properties" (Machery 2009, pp. 232–233). Moreover, following Boyd, Machery holds that a class constituting a natural kind cannot be a subset of a larger class about which the same generalisations could be formulated, because then the class in question could be subsumed by a larger—and, hence, more explanatory—class.

set of all exemplars), and all theory-like structures (but not the set of all theory-like structures). P_{CAT} is meant to stand for the cat-prototype, and the idea is that all such token prototypes are in $CONCEPT_{E_2}$ as well as all token exemplars and all token theory-like structures. In both cases, $CONCEPT_{E_x}$ cannot play an additional explanatory role in cognitive science: in the first case, because the defining property x is nothing more than a property of all the sets that play an explanatory role in cognitive science, and so does not play an explanatory role additional to the roles already played by its members. In the second case, because the complex property C_E is no more than the exclusive disjunction of the different defining properties of the explanatorily valuable kinds and so does nothing to further explain cognitive capacities. It follows that eliminativism affords CONCEPT no explanatory role, because all of the explanatory work is done on the level of representational kinds like PROTOTYPE, EXEMPLAR, and THEORY-LIKE STRUCTURE (Machery 2009, 2010).

Now consider *concept pluralism*, which also assumes that there can be different kinds of representation for one thing, but that all explanatory representations have certain properties in common:

$$CONCEPT_P = \{x \mid C_P(x) : \text{where } \forall y((P(y) \rightarrow C_P(y)) \wedge (E(y) \rightarrow C_P(y)) \wedge (T(y) \rightarrow C_P(y)))\} =$$
$$= \{p_{CAT}, e_{CAT}, t_{CAT}, p_{DOG}, e_{DOG}, t_{DOG}, \ldots\}$$

Concept pluralism reacts to the explanatory challenge by arguing that $CONCEPT_P$ does have an explanatory role, because the defining properties C_P of $CONCEPT_P$ have an explanatory role that cannot be reduced to the explanatory roles of the properties P, E, and T. The defining properties C_P of $CONCEPT_P$ are the functional properties of having (i) a logical form that allows for inferential processing; (ii) an ability to be combined; (iii) an ability to be acquired; and, finally, (iv) an ability to be stored, linked together, and retrieved by a set of memory processes (Weiskopf 2009, pp. 163–167). All of the representational kinds that are members of the set $CONCEPT_P$ are taken to possess these functional properties; i.e., the "superordinate functional roles" that all members of $CONCEPT_P$ share and that are the defining properties of $CONCEPT_P$ as a set.

According to Weiskopf (2009, p. 167):

> the existence of a set of common overarching processes and generalizations indicates that these subkinds are a more coherent and systematic object of study than their differences might otherwise lead us to think.

Thus, concept pluralism takes the kind $CONCEPT_P$ to have a unifying explanatory role, because by positing $CONCEPT_P$ we are able to explain—by reference to "functionalist considerations"—how the representational kinds posited in cognitive scientific explanation "do not belong to autonomous, disjoint systems", but rather "constitute different aspects of the human conceptual system" (Weiskopf 2009, p. 170).[8] This unifying role, however, is not played by any of the defining properties P, E, or

[8] Both concept pluralism and concept eliminativism reject "monolithic theories" of concepts that adhere to the singularity and uniformity assumptions (Weiskopf 2009, pp. 149–150). The singularity assumption

T of the more fine-grained kinds of mental representations nor can it be reduced to them.

Finally, consider *concept hybridism*. Concept hybridism concurs with concept pluralism that the kind CONCEPT has an explanatory role to play in cognitive science. However, hybridism doubts that PROTOTYPE, EXEMPLAR, and THEORY-LIKE STRUCTURE constitute disjunctive kinds. Whereas eliminativism and pluralism assume that every mental representation falls into exactly one of the three categories, hybridism assumes that a single mental representation can fall into two or even three categories at the same time. Thus, the elements of $CONCEPT_H$ do not necessarily possess only one of the three properties *P*, *E* or *T*; they may possess all three at the same time. Accordingly, $CONCEPT_H$ is a set of "integrated representations" that have PROTOTYPE-like, EXEMPLAR-like, and THEORY-LIKE STRUCTURE-like pieces of information as their parts. Thus, the elements of the set denoted by $CONCEPT_H$ are not taken to be the finer-grained representational kinds PROTOTYPE, EXEMPLAR, and THEORY-LIKE STRUCTURE, but, rather, are taken to be richly structured representations that have the potential to encode all of the pieces of information ordinarily taken to be encoded by the disjoint set of representational kinds. In the case of hybridism, therefore, our example would read:

$$CONCEPT_H = \{x \mid C_H(x)\} = \{x_{CAT}, x_{DOG}, \ldots\}$$
$$\{x \mid P(x)\}, \{x \mid E(x)\}, \{x \mid T(x)\} \subseteq CONCEPT_H$$

As Vicente and Martnez Manrique (2014, p. 61) put it:[9]

> In a nutshell, the idea [of concept hybridism] is that different structures can be regarded as constituting a common representation when they are activated concurrently, in a way that is functionally significant for the task at hand, and in patterns that remain substantially stable along different tasks related to the same category.

From the perspective of concept hybridism, the explanatory relevance of the kind CONCEPT cannot lie in an additional property that mental representations share.[10]

Footnote 8 (continued)
states that "For any category that can be conceptually represented, there is such a thing as the unique concept of that category"; and the uniformity assumption states that "All concepts belong to a single psychological kind."

[9] Since some concepts might not have all different kinds of aspects, the set of, say, prototypes might be only a subset of the set of concepts.

[10] Indeed, the defining property C_H of $CONCEPT_H$ might be nothing but the disjunctive property of the properties of PROTOTYPE, EXEMPLAR and THEORY-LIKE-STRUCTURE, such that $\forall x(C_H(x) \leftrightarrow (P(x) \vee E(x) \vee T(x)))$. The decisive difference between $CONCEPT_H$ and $CONCEPT_E$, therefore, is that the $\vee$ is to be understood as exclusive in the case of eliminativism, but as inclusive in the case of hybridism. Therefore, the two sets $CONCEPT_H$ and $CONCEPT_E$ contain very different elements. The first contains mental representations that have different functionally integrated aspects; namely, prototype-aspects, exemplar-aspects and theory-like structural aspects. The second contains prototypes, exemplars, and theory-like structures that have no common aspects. Moreover, the inclusive disjunction endorsed by concept hybridism makes it the case that while two elements of $CONCEPT_H$ need not share a single property, some may share two and some may share all three. For example, the concept ELECTRON might have only a theory-like structural aspect, while the concept DWARF might have only a prototype-aspect. However, it is likely that most representations that are the elements of $CONCEPT_H$ will have multiple aspects in common.

Rather, $CONCEPT_H$ is relevant to explanation because it makes transparent the interplay of different properties or aspects of individual mental representations in cognitive tasks according to patterns of functional integration. Thus, for the hybridist, it does not make sense to assume that PROTOTYPE, EXEMPLAR and THEORY-LIKE-STRUCTURE are mutually exclusive kinds with no overlap. Instead, a single representational token is posited—as exhibited by the members of $CONCEPT_H$—that possesses all of the different properties of the kinds PROTOTYPE, EXEMPLAR, and THEORY-LIKE-STRUCTURE at once without enforcing an internal hierarchy. It follows as a matter of course that explanations in cognitive science need not be confined to one specific aspect, but can appeal to the different aspects that are functionally integrated in members of $CONCEPT_H$. The explanatory role of $CONCEPT_H$ is to make this distinction clear.

4 Explanationism About $CONCEPT_{E,P,H}$

In this section, we will consider explanationism about the natural kind-hood of CONCEPT in the light of the different theories of CONCEPT's explanatory role introduced above. Our purpose here is to show that the explanationist approach to determining the natural kind-hood of CONCEPT stalls, because there is no consensus about what explanatory role should be afforded to the kind CONCEPT.

4.1 Explanationism About $CONCEPT_E$

According to our analysis, concept eliminativism can be read in one of two ways: either as taking the defining property x of $CONCEPT_{E_x}$ to be the property of figuring in scientific explanations of higher order cognitive capacities ($CONCEPT_{E_1}$); or as taking the defining property C_E of $CONCEPT_{E_x}$ to be the property of being a set of different explanatorily valuable kinds with (potentially) distinct defining properties ($CONCEPT_{E_2}$). On either reading, however, concept eliminativism denies that $CONCEPT_{E_x}$ has an explanatory role in cognitive science. From the perspective of the explanationist approach to determining the natural kind-hood of $CONCEPT_{E_x}$, therefore, the conclusion is clear: $CONCEPT_{E_x}$ is not a natural kind.

To illustrate why this is the case, consider the weird set that contains ELECTRON, GENE, and ANIMAL POPULATION as members. This set thus contains only kinds posited in scientific explanations:

{ELECTRON, GENE, ANIMAL POPULATION}

In this toy-example, whilst it may be the case that all of the members of the set either have the property of figuring in scientific explanation or have some defining properties of their own, it does not follow that the set {ELECTRON, GENE, ANIMAL POPULATION} is itself explanatory. Analogously for concept eliminativism, whilst it may be the case that PROTOTYPE, EXEMPLAR, THEORY-LIKE STRUCTURE all either have the property of figuring in scientific explanations

or have some defining properties of their own, it does not follow that the defining properties C_E of the set of these representations—e.g., the set $CONCEPT_{E_x}$—is itself explanatory. This holds because PROTOTYPE, EXEMPLAR, and THEORY-LIKE STRUCTURE are all taken to figure in non-overlapping scientific explanations of different cognitive processes, because their defining properties afford them different explanatory roles (cf. Machery 2009, p. 251).[11] The eliminativist, therefore, argues that the kind $CONCEPT_{E_x}$ does not yield to scientific generalisations and so is redundant in cognitive science (cf. Machery 2009, Ch. 8). In other words, because we cannot form a proposition that both features $CONCEPT_{E_x}$ and is apt to do explanatory work in cognitive science—e.g., 'all representations are *a* in virtue of being members of $CONCEPT_{E_x}$'—, the kind $CONCEPT_{E_x}$ cannot be said to have an explanatory role in cognitive science.

4.2 Explanationism About $CONCEPT_P$

In contrast to concept eliminativism, concept pluralism argues that the kind $CONCEPT_P$ does have an explanatory role to play in cognitive science, because the defining properties C_P of $CONCEPT_P$ are apt for explaining aspects of higher cognition that cannot be explained by the defining properties of the members of $CONCEPT_P$ taken individually. Thus, from the perspective of an explanationist approach to determining the natural kind-hood of $CONCEPT_P$ the conclusion is once again clear: $CONCEPT_P$ is a natural kind because it has an explanatory role to play in cognitive science, albeit to answer to "top-level [explanatory] demands" that "tend to favor unification" (Weiskopf 2009, p. 167).[12]

To make perspicuous why an explanationist approach to determining the natural kind-hood of CONCEPT will conclude that $CONCEPT_P$ is a natural kind, consider the explanatory role that $CONCEPT_P$ is afforded in virtue of possessing the defining, functional property (i): having a logical form that allows for inferential processing. According to concept pluralism, if it can be shown that all representational kinds that are member of $CONCEPT_P$ have an internal, logical structure, "then it is reasonable to suppose that there are mental processes that are sensitive to that structure, rather than to the particular concepts that are being combined in that structure" (Weiskopf 2009, p. 163). By then identifying formal inference processes that generalise over different representational kinds—e.g., the inference process that runs from 'dogs are mammals and canines' to 'dogs are mammals' and 'dogs are canines'—concept

[11] This view is consistent with the claim that PROTOTYPE, EXEMPLAR, and THEORY-LIKE STRUCTURE are domain-specific representational kinds that are suited to explain only particular domains of higher cognition. And, if this is the right way of thinking, then eliminativism is right to argue that focusing on the explanatory role of CONCEPT only distracts us from developing more accurate and empirical verified explanations of the modular—that is, the encapsulated, dissociable, automatic, neurally localized, and centrally inaccessible—operation of components of cognitive systems (Carruthers 2006, p. 62).

[12] Consider the explanatory value of the kind MAMMAL, which answers to top-level explanatory demands in the same way that, e.g., RODENTS, UNGULATES, and PRIMATES answer to bottom-level explanatory demands.

pluralists argue that the different representational kinds do, in fact, share the functional property of playing the same syntactic role in inferential thought.

Therefore, the explanatory import of $CONCEPT_P$ is justified by concept pluralists, because it is only by formulating propositions featuring $CONCEPT_P$—e.g., 'all representations have the same inferential role in thought in virtue of being members of $CONCEPT_P$'—that we can explain the inferential nature of thought *in general*, instead of having to formulate as many different explanations of inferential processing as there are representational kinds in the set $CONCEPT_P$. The same reasoning applies to concept pluralism's discussion of the functional properties (ii), (iii), and (iv) above. For example, in the case of (iii), concept pluralism argues that modes of acquisition are not sensitive to representational subkinds, because the acquisition of different representational kinds can involve abstraction over experienced exemplars and can involve the use of language and other public representational media. Therefore, the explanatory import of $CONCEPT_P$ is that it makes possible the formulation of propositions—e.g., 'all representation are acquired in processes involving x, y, z in virtue of being members of $CONCEPT_P$'—that explain the acquisition of mental states in general, instead of having to formulate as many different explanations of the acquisition of mental states as there are representational kinds in the set $CONCEPT_P$. It follows that the kind $CONCEPT_P$ has a higher-order explanatory role in cognitive science.

4.3 Explanationism About $CONCEPT_H$

Concept hybridists argue that their theory supports better—that is, more powerful—explanations of cognition, because we can explain efficiency and variability in cognitive tasks in terms of a switching between the different kinds of information encoded by members of the set $CONCEPT_H$. This follows because the members of the set $CONCEPT_H$ are thought of as being "integrated concepts" instead of as representations belonging to only one representational kind P, E, or T. So, from the perspective of an explanationist approach to determining the natural kind-hood of $CONCEPT_H$, the conclusion is yet again clear: $CONCEPT_H$ is a natural kind because it is only by formulating explanatory propositions in terms of integrated representations that we can give the best possible cognitive scientific explanations.

To make this idea explicit, consider the claim by concept hybridists that we fare better in explaining categorisation if we presuppose $CONCEPT_H$ and so posit integrated representation that have PROTOTYPE-like, EXEMPLAR-like, and THEORY-LIKE STRUCTURE-like parts. According to the hybridist, we fare better in explaining categorisation because we can appeal to the interrelated and complementary functional roles played by the integrated parts of representations in $CONCEPT_H$, depending on background factors and the task at hand (Vicente and Martnez Manrique 2014, p. 73). For instance, we can appeal to typicality effects associated with the PROTOTYPE-like part to explain why a four-legged, barking object is categorised as a dog; but, equally, we can appeal to essences associated with the THEORY-LIKE

STRUCTURE-like part to explain why we categorise 'Bobby' as a dog after hearing the sentence, 'we left Bobby in the garden to play with his chew-toy.'[13] The point, then, is that $CONCEPT_H$ must have an explanatory role in cognitive science, because without $CONCEPT_H$ we could not formulate explanatory propositions—e.g., 'categorisation involves comparing input with a dog concept, DOG_H, across various pieces of information'—that best explain the explananda of cognitive science.

In a similar vein, concept hybridists argue that their theory supports better explanations of meaning extraction. In this case, concept hybridists hold that we fare better if we posit integrated representations, because we can then provide explanations of the linguistic comprehension of lexical items in terms of our switching between different pieces of encoded information depending on context (Vicente and Martnez Manrique 2014, p. 77). For instance, we can formulate explanations that account for the processing of the lexical item 'dog' in terms of accessing the single rich concept DOG_H, even if only some parts of this concept come to be selected. In this way, the explanation can appeal to the survey and selection of the best suited information for a given task in a given context, because all information is "active and functional in meaning extraction," even if only some pieces are selected for processing to a greater or lesser extent (Vicente and Martnez Manrique 2014, p. 81). And, again, the fact that such better explanations of meaning extraction are possible only if we endorse $CONCEPT_H$ is enough for the hybridist to assert that $CONCEPT_H$ must have an explanatory role in cognitive science.

In sum, eliminativist, pluralist, and hybrid theories of concepts all have different conceptions of the kind CONCEPT: CONCEPT_E, CONCEPT_P, and CONCEPT_H respectively. Moreover, eliminativist, pluralist, and hybrid theories of CONCEPT all take their conceptions of the kind CONCEPT to support the ascription of different explanatory roles to CONCEPT. And it follows from this state of affairs that each theory provides different justifications for why CONCEPT either does (pluralism and hybridism) or does not (eliminativism) feature in propositions that best explain the explananda of cognitive science. As a result, the explanationist approach to determining the natural kind-hood of CONCEPT risks being rendered impotent, because in order to decide if CONCEPT plays a role in our best cognitive scientific explanations we first require a working consensus about which theory of CONCEPT to favour.

5 $CONCEPT_{E,P,H}$ and the Explananda of Cognitive Science

Having now introduced eliminativist, pluralist, and hybrid theories of CONCEPT and elaborated on their respective conceptions of the explanatory role of the kind CONCEPT, one point becomes apparent: that there exists a tension between these theories as to whether or not CONCEPT has an explanatory role to play in cognitive science (eliminativism vs. pluralism/hybridism) and, if it does have an explanatory role, why

[13] In cognitive science, a number of models of categorisation have already been developed that account for categorisation effects by appealing to the interplay of more than one kind of representational structure (cf. Erickson and Kruschke 1998; Anderson and Betz 2001).

it has a role to play (pluralism vs. hybridism). Here we want to argue that the reason that these two tensions obtain is as a result of eliminativist, pluralist, and hybrid theories of CONCEPT endorsing different specifications of the explananda of cognitive science.

At a superficial level, eliminativist, pluralist, and hybrid theories of CONCEPT all take cognitive science to have the same explananda: (the operations of) cognition. As a result, all three theories seem to accept that the relevant explananda associated with the positing of CONCEPT are, for instance, the kinds of category judgements and inferential processing taking place in the mind. One may suppose, therefore, that there is significant overlap between eliminativist, pluralist, and hybrid theories of CONCEPT with regards to their specifications of the explananda of cognitive science. If we dig a little deeper, however, fissures begin to appear in the descriptions of the explananda favoured by eliminativist, pluralist, and hybrid theories. For while it may be true that each view attempts to best explain particular patterns in the data; each view may also suppose that the patterns manifest in the data point to different specifications of what there is to be explained.

Consider the explanandum of category judgements as an illustration of this idea. All theories of CONCEPT will begin from the same patterns in data; typically, data that evidences particular behaviours including, but not limited to, the identification and discrimination of objects according to diagnostic features and/or properties. For the eliminativist, however, category judgements must come in a diverse number of kinds. There will be category judgements involving at least the kinds PROTOTYPE, EXEMPLAR, and THEORY-LIKE STRUCTURE; each different with respect to the salient properties identified and processed in any instance of categorising an individual c in a category C. It follows, for the eliminativist, that category judgements is not one explanandum, but several. Like the eliminativist, the pluralist will accept that category judgements come in a diverse number of kinds. For the pluralist, however, category judgements will only constitute one explanandum, because any given kind of category judgement involving any given kind of representation can be explained by our general (and entirely conceptual) ability to categorise. The hybridist concurs with the pluralist that there is only one explanandum of category judgements, but for different reasons. For the hybridist, the explanandum of category judgements does not even divide into a diverse number of different kinds of category judgements involving different representational kinds. This is the case because the hybridist takes all category judgements to involve only one representational kind: integrated and richly structured representations. Thus, the explanandum of category judgements is specified in three different ways by eliminativist, pluralist, and hybrid theories of CONCEPT even where the data to be explained—the evidence of people undertaking categorisation tasks—is the same.

The same pattern can be observed in the way eliminativist, pluralist, and hybrid theories of CONCEPT specify the explananda of inferential processing and the combination of mental representations in response to data evidencing cognitive behaviours of, e.g., reasoning and language production. In the case of the explanandum of inferential processing, for instance, eliminativists specify that there are as many different explananda of inferential processsing as there are representational kinds; pluralist specify that even if there are many different kinds of inferential processing,

all can be subsumed under the single explanandum of our general (and conceptual) ability to inferentially process; and hybridists specify that there is only one kind of inferential processing (involving integrated representations) and so there is only one explanandum of inferential processing for cognitive science to explain. The same is true of their respective specifications of the explanandum of the combination of mental representations. Here again, then, we find that eliminativist, pluralist, and hybrid theories of CONCEPT specify the explananda of cognitive science differently even when they agree on the data to be explained.

Moreover, it is not always the case that there is cross-theoretical agreement about the data to be explained. An example of this state of affairs is the relevance of chronometric data for cognitive science. The eliminativist, for instance, will likely deny that timing in task switching scenarios is an explanandum of cognitive science and so will eschew the relevance of chronometric data.[14] The reason for this likely denial is that chronometric data could speak against an eliminativist position in the following way: if chronometric data demonstrate that switching between different kinds of concepts is fast and easy—i.e., reliable—, then we would have reason to think that there is no additional cognitive mechanism that responds to context-specific processes that activate other concepts. For example, if two kinds of concepts are needed for the comprehension of a single sentence, such as in "Linda can afford to keep Bobby the dog, because chew-toys and dog licence fees are not too expensive," then hearers should need additional time to activate T_{DOG} related to the dog licence fee after having already activated P_{DOG} related to the chew-toy. If not, we have good reasons to think that theory-like structural and the prototypical pieces of information are functionally coactivated within a single dog-representation. In short, then, acknowledging chronometric data in cognitive science would *ipso facto* be an acknowledgement of the possibility of deciding if there are some processes that simultaneously involve—and, hence, unify—the objects grouped into the kind CONCEPT.

Of course, another possibility is that the eliminativist merely overlooks such chronometric data, but is ready to concede their relevance. In this case, the data would turn into empirical counter-evidence to eliminativism. Still another possibility is to find another explanation of such chronometric data that is consistent with eliminativism. Our purpose in discussing chronometric data is not to deny these possibilities, but rather to provide an example of how the interaction between a theory and its explananda could influence the interpretation of the data favoured. To make this point concrete, consider first the explananda of working memory. *Prima facie*, the data regarding working memory limitation seems to be shared by all cognitive scientists. However, we find competing explanations of working memory limitations.

[14] One will certainly not find any explicit rejection of the relevance of chronometric data in the writings of eliminativists such as Machery, but this is unsurprising. To mention such data would be to concede that such data is relevant for cognitive science, which serves only to undermine the eliminativist's position with respect to competing theories of CONCEPT. So even though the existence of chronometric data does not depend on the theory of CONCEPT one adopts, one's conception of what cognitive science aims to explain can cause such data to be irrelevant. In this sense, the data in question become "invisible" as explananda.

Miller (1956), for instance, argued that the capacity of working memory is 7 ± 2 objects (or chunks of information). Later, however, Baddeley (1992) proposed a more detailed account of working memory with different sub-systems; among them the "phonological loop" of about 2 s, the average time needed to speak about seven words in English (see also Baddeley 1996). This difference in the description of the explananda of working memory has further implications for the evaluation of each theory: while an information-theoretic approach measuring amounts of information is the obvious choice for Miller, such accounts could not explain the phenomenon as described by Baddeley; within Baddeley's model, explanations based on the articulatory apparatus are much more promising.[15] The upshot is that the data to be explained—e.g., data regarding learned responses to stimuli—are interpreted differently by the two theories: one takes it as evidence for limitations on the storage of chunks of information; the other as evidence for limitation on the storage of auditory memory traces.

Another example concerns our capacity to reason with conditionals as tested in the famous Wason Selection Task (Wason 1968; Wason and Shapiro 1971). If the conditional rule that is to be tested is formulated in an abstract way (e.g., "if there is a vowel on the one side of a card, there is an even number on the other side"), subjects perform poorly when compared to the solution that is correct according to standard sentential logic of the if-then-operator interpreted as material implication. However, if the rule is more concrete (e.g., "if a person is drinking alcohol, he must be older than 21 years"), the accuracy of subjects' reasoning improves dramatically. This is sometimes referred to as the "content effect." Given this data, one possibility is to describe the explanandum of cognitive science as the capacity of conditional reasoning, which is modulated by another factor; namely, the content of the rule. Another possibility—taken by, e.g., Cosmides and Tooby (1992)—is to deny that there is such a thing as the capacity of conditional reasoning at all, but only a capacity to deal with social rules. Thus, even though both would agree that data about the "content effect" is something that has to be taken into account in the explanation of the phenomena, they still interpret that same data in different ways: one takes it as evidence for the interaction of the capacity of conditional reasoning with some other aspect of cognition; that other takes it as evidence that the capacity of conditional reasoning should be eschewed altogether as an explanandum of cognitive science.

The problem, therefore, is that cross-theoretical agreement about the data (or the description of the data) to be explained is not always apparent. In the context of our discussion about theories of CONCEPT, the eliminativist will likely reject any interpretations of chronometric data that makes such data relevant for cognitive science. This can be seen as analogous to Baddeley's (1992) refusal to interpret the data as showing that people remember a certain amount of information; and Cosmides and Tooby's (1992) refusal to interpret that data as evidence for a general capacity of conditional reasoning that is modulated by other factors. And where there is no cross-theoretical agreement about the data to be explained or how to describe

[15] This example is only meant to be illustrative. Our intention, therefore, is not to evaluate the two theories or to decide between them.

them, we find ourselves at a loss when it comes to evaluating which of the available explanations is the best. This is also true in the context of the dispute between eliminativist, pluralist, and hybrid theories of CONCEPT. For instance, if one thinks that the explanandum of category judgements ought appeal to only one representational kind, then one has good reason for endorsing $CONCEPT_H$. But if one thinks that the explanandum of category judgements ought appeal to many representational kinds, then one has good reason for endorsing either $CONCEPT_E$ or $CONCEPT_P$; depending, that is, on one's views about the superordinate unity of those kinds. Thus, we find that one's view on the explananda of cognitive science—and, hence, on the interpretation and (ir)relevance of certain bodies of data (or their description)—bias one towards a certain theory of CONCEPT and towards a certain view of CONCEPT's explanatory role.

6 Capacities and Effects

Now, one may think that our conclusion in the previous section is too quick, because we ourselves have not provided a specification of the explananda of cognitive science. Plausibly, then, one may think that it is at least possible to specify the explananda of cognitive science in such a way that would undermine the apparent disagreements between eliminativist, pluralist, and hybrid theories of CONCEPT. For example, one may think it is possible to follow Cummins (2000, p. 120) and argue that the explananda of cognitive science should be divided into primary and undiscovered *capacities*—e.g., "to see depth, to learn and speak a language, to plan, to predict the future, to empathize, to fathom the mental states of others" (Cummins 2000, pp. 124–125)—and secondary and discovered *effects*—e.g., well confirmed regularities that can be specified as laws *in situ* that "restate the phenomenon in more general terms" (Cummins 2000, p. 120).[16] Working from this premise, one could argue that:

> the explanation of incidental effects [...] have little interest in their own right: no one would construct a theory just to explain them. But their successful explanation can often be crucial to the assessment of theories or models designed to explain the core capacities that are the primary targets of psychological inquiry (Cummins 2000, p. 128).

Accordingly, one could submit the following counter-argument to our claim: the disagreement between eliminativist, pluralist, and hybrid theories of CONCEPT is about the effects identified by cognitive science and not about the capacities that cognitive science aims to explain. If this counter-argument were right, then the dispute between these theories would not be about the explananda of cognitive science *tout court*, but would be about fine-grained explanatory issues found at the level of

[16] A good example of an effect would be the McGurk effect, which can be paraphrased as a law that states that one will have the illusion of hearing a particular sound when the auditory component of another sound is paired with the visual component of yet another sound.

effects; for instance, the speed of categorisation and shifts in categorisation when different aspects of the stimuli are emphasised (Ahn and Kim 2000; Ahn and Dennis 2001). In Cummins' terms, the dispute would be about "what happens" and not "why or how"; and this would lead us to be more optimistic about finding a specification of the explananda that could be shared by all theories of CONCEPT.

This counter-argument, however, fails to appreciate the difficulty in differentiating between capacities and effects when we factor in the contradictory viewpoints endorsed by the different theories of CONCEPT. For it is clear that specifying the explananda of cognitive science in terms of both capacities and effects is highly non-trivial. Cummins (2000, p. 127) himself states that "it can be a matter of substantive controversy whether we are looking at an exercise of a capacity or an incidental effect." This controversy is heightened in the case of the debate between different theories of CONCEPT, because the question of how to draw the line between capacities and effects cannot be conveniently segregated from a deeper question about what the capacities are in the first place.[17] For example, one could argue—in accord with the pluralist—that differences in categorisation judgements involving different kinds of representations are merely effects incidental to the exercise of a single capacity to categorise. But, equally, one could take the eliminativist view that there are as many different capacities to categorise as there are representational kinds operative in cognition. And this highlights an important point; namely, that there will be no agreement between eliminativists, pluralists, and hybridists about how to enact a functional analysis that delivers a demarcation between capacities and effects. And thus there will be no agreement about which explanations best explain either capacities or effects, or about the structure of the system giving rise to both capacities and effects.[18]

The same point can be made against those who argue that we do not consider in enough detail the explanatory targets of working cognitive scientists. For example, those who insist that working cognitive scientists could never get on board with CONCEPT eliminativism, because the kind CONCEPT is to them an indispensable explanatory tool. Keil (2010, p. Keil), for instance, argues that there will be "a strong tendency to resist" the claim that there are "an indefinitely large number" of representations operative in cognition (e.g., $p_{DOG}, e_{DOG}, t_{DOG}, p_{CAT}, e_{CAT}, t_{CAT} \ldots$). Underlying this claim is the worry articulated by Hampton (Hampton 2010, p. 212) that:

> the term "concept" is needed as part of an account of the many situations in which PET systems [(e.g., PROTOTYPES, EXEMPLARS, and THEORY-LIKE STRUCTURES representations)] interact. How does one discuss concept combination, including the formation of composite prototypes, the importing of exemplar knowledge, and the coherence checking of the result through background theory, if one cannot have the integrative term "concept" to specify just what it is that is

[17] This point connects to our discussion of the putative capacities of working memory and conditional reasoning in the previous section.

[18] Note that we do not want to take a stand on how we should specify the explananda of cognition. Rather, we only want to show that different specifications are possible but will be mutually contradictory. This, in turn, problematises the appeal to cognitive science by theories of CONCEPT.

being combined. The combination occurs at the concept level, and the description of the processes involved then requires elaboration in terms of the PET systems.

The counter-argument, therefore, is that given the state of cognitive scientific research there are some explananda—e.g., explananda that require cross-representational processing such as "concept combination"—that demand that CONCEPT be afforded an explanatory role in cognitive science. The problem, however, is that one need not endorse the claim that putative explananda involving cross-representational processing are part of the explanatory remit of cognitive science. Instead, one may think that the composition of PROTOTYPES is distinct from the composition of EXEMPLARS and THEORY-LIKE STRUCTURES; the use of EXEMPLAR knowledge is distinct from the use of PROTOTYPE and THEORY-LIKE STRUCTURE knowledge; and that coherence checking is limited to one representation kind at a time. Thus, in terms of Cummins' distinction, one may hold that the capacities associated with each kind of representation are distinct and that the specification of an effect of cross-representational processing fails to pick out a regular behavioural patterns characteristic of the structure of cognition. This strictly modular view of cognitive structure may strike some as unappealing, but it will dovetail with the specification of the explananda favoured by the eliminativist and with the eliminativist's view on CONCEPT's explanatory role.[19]

The upshot is that the appeal to the working explanatory interests of cognitive scientists underdetermines the specification of the explananda. For while it is true that many cognitive scientists have been willing to characterise behavioural patterns as characteristic of a particular kind of structure responsible for cross-representational processing, it is also true that all cognitive scientists need not characterise the same behavioural patterns in the same way. For instance, one may characterise an infant's switch from PROTOTYPE-based categorisation judgements to THEORY-LIKE STRUCTURE-based categorisation judgements in terms of a kind of structure responsible for cross-representational processing (Keil 1989). But, equally, one may characterise the same switch as a binary change in the operation of two, distinct capacities: the capacity to categorise using PROTOTYPES and the capacity to categorise using THEORY-LIKE STRUCTURES. The point, then, is that one cannot assume *ex ante* what the capacity or capacities for conceptual change consists in, because it is possible that the switch in development from categorising with PROTOTYPES to categorising with THEORY-LIKE STRUCTURES is a mere incidental effect. Our argument is that the view one takes on these matters will cohere with the theory of CONCEPT and of CONCEPT's explanatory role one favours.

[19] It is worth making explicit at this point that we do not want to argue that there are *no* reasons to accept one or another theory of CONCEPT. Of course, one could find any number of reasons; for example, reasons concerned with putative theoretical virtues such as *beauty*, *simplicity*, and *coherency* (cf. Keas 2018, for a good summary of such virtues); or sociological reasons concerned with one's experience with and preference for distinct explanatory tools or one's institutional embedding. Our only argument, then, is that the explanatory success of theoretical terms like CONCEPT cannot be determined independent of a theory, and thus there is no out-of-theory reason to accept this or that ontological claim about the existence of such things as concepts.

To sum up, we do not argue that a unification of cognitive science is, in principle, impossible. Rather, we argue that we do not currently have a unified specification of what cognitive science aims to explain, as evidenced by the fact that different theories of CONCEPT take cognitive science to be targeting different explananda. Although it is clear that are overlaps in what different theories of CONCEPT take cognitive science to be in the business of explaining, there is also enough disagreement to undermine the search for a definitive account of CONCEPT's explanatory role. Thus, to make progress in this regard we would first have to arrive at a unified specification of the explananda of cognitive science. In the next section, however, we will argue that the presence of the divergent theories of CONCEPT makes it doubtful that any unified specification could be attained, which serves to undermine the explanationist approach to determining the natural kind-hood of CONCEPT.

7 CONCEPT, Explananda, and Explanationism

In Sect. 4, we argued that eliminativist, pluralist, and hybrid theories of CONCEPT disagree about the explanatory role of CONCEPT in virtue of endorsing three different interpretations of the concept of concept; that is, in virtue of endorsing $\text{CONCEPT}_{\text{E}}$, $\text{CONCEPT}_{\text{P}}$, and $\text{CONCEPT}_{\text{H}}$ respectively. And in section five and six, we argued that the tension between eliminativist, pluralist, and hybrid theories of CONCEPT is due to each theory endorsing different specifications of the explananda of cognitive science. The remaining question, however, is how this undermines an explanationist approach to determining the natural kind-hood of CONCEPT.

Even given what we have said above, an advocate of an explanationist approach to natural kind-hood determination may hold that the natural kind-hood of CONCEPT can be determined by evaluating whether or not it participates in the *best* cognitive scientific explanations—it is just that we do not know what these explanations are yet or, crucially, what these explanations purport to explain. As we have argued above, however, the tension between the three theories about the explanatory role of CONCEPT arises from different interpretations of what the best explanations of cognitive science set out to explain.[20] Therefore, we contend that in order for the explanationist approach to determining the natural kind-hood of CONCEPT to be viable, we must first decide what the explananda of cognitive science are; which—according to our argument—amounts to the same thing as deciding between eliminativist, pluralist, and hybrid theories of CONCEPT. Thus, there is circle built into explanationism that prevents it from being a viable method to determine whether CONCEPT is a natural kind.

In principle, we can always disagree about the best cognitive scientific explanations for a given explanandum. However, as we demonstrated in Sects. 5 and 6, the dispute between eliminativist, pluralist, and hybrid theories of CONCEPT is at an even deeper level, for they do not even agree about the explananda to be explained by

[20] Note that all three theories aim to account for why cognitive science does or does not need to employ CONCEPT in explanation; and since we do not have reason to assume that they speak about three different cognitive sciences, there must be some reason for their divergent perspectives in this regard.

cognitive science in the first place. For example, in one case the kind $CONCEPT_P$ is taken to explain the unity of all category judgements (concept pluralism); but, in another case, the kind $CONCEPT_E$ is taken to explain nothing at all such that the unity of all category judgements is explicitly rejected. In this way, one theory cites explananda that are not cited or even countenanced by the other. Thus, we cannot even say that in all cases one theory of CONCEPT rejects the cognitive scientific explanations favoured by another as bad or superfluous explanations. Instead, we must say that in at least some cases one theory rejects the *interpretation* of cognitive scientific explanations favoured by another as bad or superfluous interpretations. The point, then, is that because eliminativist, pluralist, and hybrid theories of CONCEPT do not agree on the explananda for which we seek cognitive scientific explanations, they cannot agree on what counts as the best cognitive scientific explanations.

It is clear that all sciences must begin with a specification of their explananda. For example, the phenomenon of lightning was specified as an explanandum of physics, and consequently physics has formulated an explanation for this phenomenon. But we must also keep in mind that the possibility of progress depends upon there being better specifications of the explananda, which serve to restrict the number of acceptable explanations that the science can formulate. It follows that the act of specifying the explananda of a certain science is not independent of the progress of that science—in fact, the progress of a science feeds into a process by which better specifications of the explananda are made and new explananda are brought into the purview of explanation.[21] For sure, it is not an easy question why certain phenomena belong to the subject matter of a certain science. Moreover, during the development of a science, the range of phenomena belonging to the subject matter of this science is liable to change. Not only are new phenomena identified (e.g., quantum effects), but known phenomena might 'change sides' (e.g., some 'chemical' facts about the reactivity of substances turned out to be better explainable by atomic physics). We are thus confronted with a situation that is sometimes called a "virtuous circle," where the explanatory goals of a science (its explananda) and the explanations that the science provides are mutually constitutive. Schematically:

1. The explananda of a science are (partly) determined by specifying them.
2. The terms used to specify the explananda are determined by the theory (i.e., they are theoretical terms).
3. As the theory develops, new explanations are found.
4. As new explanations are found, new terms are introduced and existing terms are refined.
5. With new and refined terms, the explananda of the science change.[22]

[21] It is clear that a science evolves and develops as it tries to improve upon its best current explanations. So-called "theoretical terms" are introduced into science for the very reason of making something explainable (i.e., by abduction or inference to the best explanation). By parsing scientific theories in terms of, e.g., Ramsey-sentences or partial structures, such terms and their defining place within a theory can be made explicit (cf. Andreas 2017; Lewis 1970; Van Fraassen 1980).

[22] To be sure: the phenomena to be explained do not change, but their descriptions do, which essentially involve scientific terms. Since the explananda (as understood here) are not the phenomena themselves but the specific descriptions of the phenomena, the explananda change. E.g., the explanandum of why all objects fall with the same speed in a vacuum was not available before Galileo.

In the current debate about the natural kind-hood of CONCEPT, a positive portrayal of the virtuous circle has been presupposed, where explananda-specification takes us first to viable explanations, then we move from viable explanations to refined specifications of the explananda, and finally from refined specifications of the explananda back again to better explanations. On this picture, the explanationist approach to determining the natural kind-hood of CONCEPT can seem to make some sense: CONCEPT is a natural kind only insofar as it features in the better explanations arrived at following the interchange between explananda-specification and cognitive scientific explanation. This view, however, is brought into question when we recognise that any answer to the question of what constitutes a 'better' cognitive scientific explanations must presuppose an answer to the question of what explananda are to be explained by cognitive science in the first place. Thus, we are forced to accept that a specification of the explananda of cognitive science is never theoretically innocent, because it serves to constrain the process of formulating 'better' explanantions and 'better' explananda-specifications further down the line. It follows that one's view of what counts as the best explanation will be influenced by one's view of the explananda in need of explanation.

To decide between theories of CONCEPT, it seems therefore that we would first have to find a way to settle the explananda of cognitive science. But this cannot be easily achieved. To illustrate this point, consider the following two explananda: (a) a stick half under water that looks bent even though it is not; and (b) two lines of the same length, one with inward pointing arrow heads, the other with outward pointing arrow heads, which look like they are of different length even though they are not (the "Müller–Lyer–Illusion"). The first explanandum is one of optics, the second is one of psychology of perception. Accordingly, the first is easily explained by the laws of optics, whereas the second is not; to explain the second phenomenon, we need to appeal to basic psychological principles of perception that are not related to the laws of optics. However, why (a) and (b) belong to the subject matter of different disciplines is not *prima facie* obvious, and we doubt that there could be a specification of the explananda that would make the difference clear, unless that specification already presupposes the difference between optics and psychology. For example, one could try to specify that (a) is an explanandum belonging to the subject matter of optics and (b) an explanandum belonging to psychology by arguing that everything in front of the retina is optics; and since (the image of) the stick is bent on the retina but (the images of) the two lines are not of different sizes on the retina, the first would be specified as an optical phenomenon and the second not. However, this specification already presupposes that optics is confined to certain visual phenomena and psychology to the processing of visual phenomena, which presupposes a certain understanding of the disciplines and their subject matter, which then biases our specifications of the explanandum itself.[23]

[23] One can see clearly here how finding better specifications of the explananda is part of the remit of science. For example, as soon as the the "Müller–Lyer–Illusion" is identified as a psychological explanandum, psychology will find better specifications of the phenomenon giving rise to the explanandum; that is, that it is a phenomenon made manifest by a default heuristic in the visual system that processes the configuration of angled lines so as to optimise judgements about depth and distance (Gregory 1966).

As with our example of the specification of explananda (a) and (b), eliminativist, pluralist, and hybrid theories of CONCEPT endorse specifications of the explananda of cognitive science that dovetail with their understanding of the discipline and its subject matter. This point has been made at length in Sect. 5. With this in mind, we can draw one final conclusion for our discussion of the different theories of CONCEPT: we cannot decide which specification of the explananda of cognitive science to endorse by simply looking at the kinds of explanations formulated in cognitive science. The reason for this state of affairs is because every specification of the explananda will be validated by those explanations that are taken to explain the explananda specified. That is, by those explanations that can be interpreted as explaining, say, the unity of cognitive systems (concept pluralism), the modularity of cognitive systems (concept eliminativism), or the functional integration of the representations operative in cognitive systems (concept hybridism). And this bias runs both ways, because no comparison of cognitive scientific explanations will be possible where there is disagreement about what it is that cognitive science ought to be in the business of explaining. We thus seem to lose the basis for a comparison of theories of CONCEPT that is not *ad hoc*, since different theories presuppose different specifications of the explananda of cognitive science, and so each has reasons to find different cognitive scientific explanations more or less successful.[24] If we couple this argument with our claim in Sects. 3 and 4 that different theories of concept afford CONCEPT different explanatory roles, then it is evident that an explanationist approach to determining the natural kind-hood of CONCEPT cannot be made to work. This follows because we cannot hope to decide between the different theories of CONCEPT's explanatory role (e.g., $CONCEPT_{E_x}$, $CONCEPT_P$, and $CONCEPT_H$ respectively) when we cannot even agree about what cognitive science aims to explain.[25]

8 Outlook for Discussions About CONCEPT

Our conclusion, if correct, appears to leave the naturalistically-inclined philosopher of cognitive science in a difficult spot. For it seems that the explanations formulated in cognitive science can no longer be taken as a reliable guide to the natural

[24] Our argument here has certain parallels with the idea of "experimenter's regress" put forward by Collins (1981) and Collins (1992) in his discussion of Joseph Weber's apparatus for gravitational wave detection. According to Collins, there is a circle between judgements of the validity of a measurement device and judgement of the validity of a measurement result. More specifically, he argues that "we don't know if we have built a good detector until we have tried it and obtained the correct outcome! But we don't know what the correct outcome is until...and so on *ad infinitum*" (Collins 1992, p. 84). Thus, a regress obtains when "scientists try to justify their judgements about a given outcome or about the quality of their data" (Feest 2016, p. 35). We accept that our claim that there is no way to compare explananda across theoretical contexts is analogous to Collins' claim that there is no way to adjudicate disagreements over whether a particular empirical test has captured a certain phenomenon. However, we will set aside further discussion of the relation between the two positions—and of the viability of Collins' claims (cf. Franklin 1999, for criticism of Collins' position)—due to lack of space.

[25] To prevent misunderstanding: our argument does not exclude a future specification of the explananda of cognitive science that incorporates all the different explananda of eliminativism, pluralism, and hybridism, and so is able to offer a unified account with respect to these three theories of CONCEPT. However, this new overarching specification of the explananda would be just another specification, possibly

kind-hood of CONCEPT. One may suppose, therefore, that the kind CONCEPT—and perhaps other similar kinds featuring in cognitive scientific explanation—cannot be said to be natural kinds at all. This line of reasoning, however, is much too quick. For whilst it may be true that the kind CONCEPT ought be thought of as a mind-dependent, social kind in Hacking's (1995; 1999) sense; it does not follow that CONCEPT cannot also be a natural kind (Hacking 1995, 1999; Khalidi 2009). This is the case because even if the classification of CONCEPT is interactive and can change in response to our attitudes towards cognitive scientific explanations, this does not make the CONCEPT ontologically subjective. The crucial point here has been put clearly by Khalidi (2015) as follows:

> Mind-dependence is a red herring when it comes to ontological objectivity. There are various phenomena that depend on the human mind (both causally and constitutively) yet are not non-real, at least not in the same sense as fictional entities. Still, isn't there a sense in which all social kinds are ontologically subjective, as Searle claims? Doesn't the fact that they would not have existed without the existence of human minds render them ontologically different from other kinds? Some perspective on these questions may be gained by reflecting further on the analogy between mind and life. Consider biological kinds like tiger, larva, and metabolism. It is safe to say that these biological kinds are life-dependent, in the sense that they would not have existed without life. But that does not seem to impugn their ontological objectivity, and nor should the mind-dependence of social (and psychological) kinds (Khalidi 2015, 111).

The central message of this paper, then, is not that CONCEPT is not a natural kind. Rather, we hope to have shown that any approach to determining the natural kind-hood of CONCEPT must pay attention to the mind-dependence of the kind CONCEPT as a tool in the ongoing, non-monotonic practice of explanation-giving and explananda-specification in cognitive science. The explanationist approach to determining the natural kind-hood of CONCEPT fails to recognise this point, because even where explanationism accepts the plasticity of explanations and explanatory methods, it must assume that the best explanations—and, hence, the 'correct' specifications of the explananda—can always be agreed upon (Poston 2016). But since there are no theory-neutral standards of epistemic justification by which to compare explanations and explananda-specification, this assumption is misguided (Appley and Stoutenburg 2017; Stoutenburg 2015). As a result, there can be no straightforward explanationist determination of the natural kind-hood of CONCEPT.

Having made this point, we need not go as far as to say that discussion of CONCEPT's natural kind-hood should be divorced from cognitive science altogether. We must, however, pay attention to the fact that progress is only possible when there is the possibility of tractable disagreement about the best cognitive scientific explanations and

Footnote 25 (continued)

rivalling a second future specification, which could then go on to rival a third future specification, and so on *ad infinitum*. So, our argument in this paper would apply equally to all such hypothetical future specification of the explananda of cognitive science.

explananda-specifications. In this way, we must recognise that the results of cognitive science can only come to bear on discussions about the natural kind-hood of CONCEPT when we are able, in principle, to come to an agreement about what constitutes the best cognitive scientific explanations and explananda-specifications. We are not convinced that any such agreement is possible. But, if it is, then it is most likely to be found in the case of those explanations that account for explananda specified in accordance with our shared, pre-theoretic understanding of the world; that is, the *sui generis* explananda specified by attending to our common experiences. If there could be a set of best explanations of these explananda, then we could implement a revised explanationist approach to natural kind-hood determination, whereby we determine natural kind-hood by identifying those kinds that play a role in explanations that best explain pre-reflectively specified explanada. But the promise of using such a revised explanationist approach to determine the natural kind-hood of CONCEPT depends, counter-intuitively, on a pre-theoretic specification of what there is for cognitive science to explain.

Acknowledgements We would like to thank both anonymous reviewers for their helpful recommendations and advice about how the paper could be improved. Thanks also to participants of the *Concepts and Explanation* conference in Düsseldorf for their insightful and constructive comments and critique. Finally, thanks to all our colleagues in the CRC and the philosophy department at Heinrich-Heine-University for their support. This work was funded by the DFG (German Research Foundation) as part of the Collaborative Research Centre 991: The Structure of Representations in Language, Cognition, and Science.

References

Ahn, W., & Dennis, M. J. (2001). Dissociation between categorization and similarity judgement: Differential effect of causal status on feature weights. In U. Hahn, & M. Ramscar (Eds.), *Similarity and categorization* (pp. 87–107).

Ahn, W., & Kim, N. S. (2000). The causal status effect in categorization. *Psychology of Learning and Motivation: Advances in Research and Theory*, *40*, 23.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review*, *8*, 629–647.

Andreas, H. (2017). Theoretical terms in science. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy (Spring 2018 Edition)*. http://plato.stanford.edu/archives/sum2013/entries/theoretical-terms-science/. Accessed July 2018.

Appley, B. C., & Stoutenburg, G. (2017). Two new objections to explanationism. *Synthese*, *194*(8), 3069–3084.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 5–28.

Bird, A., & Tobin, E. (2017). Natural kinds. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)*. https://plato.stanford.edu/archives/spr2018/entries/natural-kinds/. Accessed July 2018.

Bloch-Mullins, C. L. (2018). Bridging the gap between similarity and causality: An integrated approach to concepts. *British Journal for the Philosophy of Science*, *69*(3), 605–632.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical studies*, *61*(1–2), 127–148.
Boyd, R. (1999). Kinds, complexity and multiple realization. *Philosophical Studies*, *95*(1–2), 67–98.
Byerly, T. R. (2013). Explanationism and justified beliefs about the future. *Erkenntnis*, *78*(1), 229–243.
Carey, S. (1985). *Conceptual change in childhood*. Cambridge: Cambridge University Press.
Carruthers, P. (2006). *The architecture of the mind*. Oxford: Oxford University Press.
Collins, H. (1981). Son of seven sexes: The social destruction of a physical phenomenon. *Social Studies of Science*, *11*(1), 33–62.
Collins, H. (1992). *Changing order: Replication and induction in scientific practice*. Chicago: University of Chicago Press.
Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, *163*, 163–228.
Cummins, R. C. (2000). "How does it work?" versus" what are the laws?" Two conceptions of psychological explanation. In F. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). MIT Press.
Ellis, B. (2001). *Scientific essentialism*. Cambridge: Cambridge University Press.
Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology*, *127*, 107–140.
Feest, U. (2016). The experimenters' regress reconsidered: Replication, tacit knowledge, and the dynamics of knowledge generation. *Studies in History and Philosophy of Science Part A*, *58*, 34–45.
Fodor, J. A. (1994). Concepts: A potboiler. *Cognition*, *50*, 95–113.
Franklin, A. (1999). How to avoid the experimenters' regress. In A. Franklin (Ed.), *Can that be right?* (pp. 13–38). Berlin: Springer.
Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
Gregory, R. (1966). *Eye and brain*. New York: McGraw-Hill.
Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary approach*. Oxford: Oxford University Press.
Hacking, I. (1999). *The social construction of what?*. Cambridge, MA: Harvard University Press.
Hampton, J. A. (2010). Concept talk cannot be avoided. *Behavioral and Brain Sciences*, *33*(2-3), 212–213.
Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, *195*(6), 2761–2793.
Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
Keil, F. C. (2010). Hybrid vigor and conceptual structure. *Behavioral and Brain Sciences*, *33*(2-3), 215–216.
Ketland, J. (2004). Empirical adequacy and ramsification. *The British Journal for the Philosophy of Science*, *55*(2), 287–300.
Khalidi, M. A. (2009). Interactive kinds. *The British Journal for the Philosophy of Science*, *61*(2), 335–360.
Khalidi, M. A. (2015). Three kinds of social kinds. *Philosophy and Phenomenological Research*, *90*(1), 96–112.
Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford: Oxford University Press.
Lakoff, G. (1987). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 63–100). Cambridge: Cambridge University Press.
Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, *67*(13), 427–446.
Machery, E. (2009). *Doing without concepts*. New York: Oxford University Press.
Machery, E. (2010). Precis of doing without concepts. *Behavioral and Brain Sciences*, *33*, 195244.
Machery, E., & Seppälä, S. (2011). Against hybrid theories of concepts. *Anthropology and Philosophy*, *10*, 99–126.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.
Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.
Poston, T. (2016). Explanationist plasticity and the problem of the criterion. *Philosophical Papers*, *40*(3), 395–419.
Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge: MIT Press.
Psillos, S. (2005). Scientific realism and metaphysics. *Ratio*, *18*(4), 385–404.
Ramsey, F. P. (1931). Theories. In *The Foundations of Mathematics* (pp. 212–236). London: Routledge.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141–1159.
Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.
Saatsi, J. (2017). Explanation and explanationism in science and metaphysics. In M. Slater & Z. Yudell (Eds.), *Metaphysics and the philosophy of science: New essays* (pp. 163–192). Oxford: Oxford University Press.
Stoutenburg, G. (2015). Best explanationism and justification for beliefs about the future. *Episteme*, *12*(4), 429–437.
Suppes, P. (1967). What is a scientific theory? In S. Morgenbesser (Ed.), *Philosophy of science today* (pp. 55–67). New York: Basic Books.
Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford: CSLI Publications.
Suppes, P. (2008). Abstract entities. In T. Sider, J. Hawthorne, & D. W. Zimmerman (Eds.), *Contemporary debates in metaphysics* (pp. 11–31). Oxford: Blackwell.
Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.
Vicente, A., & Martnez Manrique, F. (2014). The big concepts paper: A defence of hybridism. *The British Journal for the Philosophy of Science*, *67*(1), 59–88.
Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.
Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology*, *23*(1), 63–71.
Weiskopf, D. A. (2009). The plurality of concepts. *Synthese*, *169*, 145–173.