European Journal for Philosophy of Science (2021) 11: 19
https://doi.org/10.1007/s13194-020-00344-3

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Natural kind terms again

Panu Raatikainen[1]

Received: 17 July 2020 / Accepted: 15 December 2020 / Published online: 5 January 2021

Abstract
The new externalist picture of natural kind terms due to Kripke, Putnam, and others has become quite popular in philosophy. Many philosophers of science have remained sceptical. Häggqvist and Wikforss have recently criticised this view severely. They contend it depends essentially on a micro-essentialist view of natural kinds that is widely rejected among philosophers of science, and that a scientifically reasonable metaphysics entails the resurrection of some version of descriptivism. It is argued in this paper that the situation is not quite as dark for the new theory of reference as many critics suggest. There are several distinct questions here which should not be conflated and ought to be dealt with one by one. Descriptivism remains arguably problematic.

Keywords Natural kinds · Reference · Descriptivism · Essentialism

1 Introduction

In the last few decades, the new externalist picture of natural kind terms due to Kripke, Putnam, and others has become quite popular in philosophy. This new theory of reference (NTR), as it is often called, has contested more traditional descriptivist and internalist views of meaning and reference, in particular in the case of proper names and natural kind terms.

The reactions of philosophers of science to this have been mixed: some have been rather enthusiastic, whereas the attitudes of others have been mainly critical. In a recent article, Sören Häggqvist and Åsa Wikforss (henceforth H&W) severely criticise NTR as applied to natural kind terms.[1] They contend that it depends essentially on a view of natural kinds that is 'widely rejected among philosophers of science' (p. 912). According to H&W, a 'scientifically reasonable metaphysics' entails the 'resurrection of some version of descriptivism' (p. 913). The article by H&W is only the latest in a long series

[1]All page references, unless otherwise stated, are to Häggqvist and Wikforss 2018.

✉ Panu Raatikainen
panu.raatikainen@tuni.fi; https://philpeople.org/profiles/panu–raatikainen

[1] Philosophy, Tampere University, Pinni B4147, FIN-33014 Tampere, Finland

of attacks against NTR from the perspective of the philosophy of science. Zemach (1976), Mellor (1977), Dupré (1981), Hacking (1991, 2007), LaPorte (1996, 2004), and Needham (2000, 2002, 2011), for example, have put forward critical considerations not dissimilar to those of H&W. In addition, Lewis has been a quite influential counterbalance in his resistance to NTR (Lewis 1994). H&W sum up much of such previous critical arguments and countermoves, though they certainly also add their own ingredients to the mix. Consequently, a thorough analysis of the arguments of H&W may simultaneously be significant for this critical literature more generally.

There is no question that H&W and other critics raise several important issues and make many apt observations. Things indeed tend to be much more complex in reality than philosophers' pet examples may suggest. Reminders of just how complicated things are in the real world can be important, but the philosophical consequences of such facts must not be inflated. I shall argue that the situation is not quite as dark for the externalist view or NTR as many opponents suggest. There are several distinct questions here which must not be conflated and should be dealt with one by one. On closer scrutiny, it is not entirely clear what exactly it is that H&W, for example, ultimately achieve.[2] In particular, I shall argue that the reports of descriptivism's resurrection are greatly exaggerated.

What more precisely is the view at issue here? For many critics of NTR, including H&W, this is not completely unambiguous. First, H&W say they are criticising externalism. A moment later they say that they are defending a form of descriptivism. In what follows, they are instead attacking micro-essentialism. Similar slides from one issue to another are common in the critical literature. Though these views (or the directly opposing views) have been historically related, they are distinct issues and should be considered separately. H&W contrast externalism with descriptivism. This is a bit confusing: descriptivism should be contrasted with 'the nondescription theory'.[3] Externalism is opposed to internalism, not descriptivism. Internalism is the view that meaning is always determined by a language user's narrow mental states.[4] On the one hand, internalism is a more general view than descriptivism, since it does not require that the relevant narrow mental states are related to associated descriptions. On the other hand, associating certain kinds of descriptions with an expression may result in a mental state that is *not* narrow and hence does not support internalism.[5] These differences become relevant in some contexts. Finally, neither the nondescription theory nor externalism is necessarily committed to straightforward and overarching

[2] Also Hoefer and Martí 2019 have put forward a critique of H&W; Hoefer and Martí 2020 is relevant too. I am largely in agreement with them, but in the present paper my focus is on somewhat different issues than in those papers. I think their papers and the present one nicely complement each other and together make a strong case in favour of NTR in this debate.

[3] I have borrowed this label from Devitt 1989. Here we can take it to mean the negation of the view that names and kind terms are synonymous with (clusters of) descriptions which language users analytically associate with these words and that these meaning-expressing (clusters of) descriptions determine the extensions of the words.

[4] How exactly 'narrow' should be defined has turned out to be more difficult than Putnam (1975) initially assumed; see Burge 1982; Williamson 2006; Yli-Vakkuri and Hawthorne 2018. Roughly, a state is narrow if it depends only on the total internal qualitative state of the subject; an internally identical Doppelgänger has the same narrow states even in different environments. In any case, some such narrowness condition is necessary, for otherwise 'internalism' threatens to become trivially true, and the whole issue philosophically uninteresting; see e.g. Devitt 1990.

[5] Cf. Section 6 below.

micro-essentialism of any kind, or so shall I argue. Be that as it may, when the details do not matter so much, I shall simply talk about 'the new theory of reference' or 'NTR' (referring vaguely to the cluster of views associated with Kripke, Putnam, and others) and the opposition to it.

My primary goal is not, however, to attack H&W for the sake of it. Rather, I shall use their critical discussion as a baseline, and my principal aim is to clarify the whole area and what really is central to NTR and what is not, and to put forward certain positive ideas about natural kind terms (or at least about a certain sort of natural kind terms) and their meaning and reference.

2 Theories of reference revisited

It is important to put critical discussions of NTR in the right context in the theories of meaning and reference, and to keep the big picture clear in our mind. Only then we can properly evaluate their relevance. Although the story is familiar, let me briefly summarise it. It is common to begin surveys of theories of meaning and reference with 'the direct reference theory' regularly ascribed to Mill. This is the simple view according to which the *meaning* of a name is simply the object it denotes. In the case of general terms, the analogous view holds that meaning is simply extension, that is, the set of entities that the term applies to; this latter view has sometimes been called 'extensionalism' (see e.g. Braun 2006).

Such views encounter well-known difficulties in 'Frege's puzzles' (e.g. pairs of expressions with the same referent or extension but which intuitively have a different meaning). Consequently, descriptivism, put forward as a more plausible alternative, proposes that there must be *more* to the *meaning* of a referring expression than the entity named or the extension – namely the *descriptive content* of the expression (what the associated description, or the cluster of descriptions, expresses).[6] In a parallel way, H&W explain descriptivism (for kind terms) as follows: 'the meanings of kind terms are given by a set of associated descriptions (typically relating to easily available, observable features), determining the extensions of the terms' (pp. 911–12). They thus apparently agree that descriptivism is essentially a theory of *meaning*.[7]

Kripke (1972/1980), Donnellan (1970), Devitt (1981), and others in turn have argued against descriptivism, and Putnam (1975) and Burge (1979) have argued against internalism. Their critical attacks share in common *arguments based on the ignorance and error* of ordinary language users (Putnam's famous Twin Earth thought experiment can be viewed as a powerful argument from ignorance). Kripke and especially Putnam suggested that whether something belongs to the extension of a natural kind term or not

[6] In the case of the modern cluster theory version of descriptivism, it is required that all language users competent with the expression associate with it at least a sufficiently large part of the cluster of descriptions (even if different language users may perhaps associate with somewhat different parts of the cluster).

[7] Kripke also considers briefly a weaker possible version of descriptivism, which says nothing about *meaning* but only contends that the reference of an expression is determined by the description associated with it (Kripke 1972/1980, p. 31–32). Kripke adds that 'some of the attractiveness of the theory [descriptivism] is lost if it isn't supposed to give the meaning of the name'; this is because it is not clear it can still solve Frege's puzzles which concern meaning (Kripke 1972/1980, p. 33). All in all, it is descriptivism understood as a theory of meaning that is a well-motivated, natural, and unified whole, as well as the main target of NTR. Clearly also H&W understand descriptivism in this standard way as a theory of meaning. This becomes relevant below.

may depend in part on some (at the given time) unknown underlying features of the original samples and is not determined solely by some readily observable characteristics well known and associated with the term by all the competent language users. Kripke presented in addition certain modal considerations against descriptivism, including the argument from rigidity.[8] Kripke also put forward a rough sketch of an alternative view, the causal-historical picture of reference. Others have largely adopted that picture.

As H&W also note, Kripke's causal-historical picture falls into two parts. First, there is the initial *introduction* of a referring expression to the language, a '*baptism*', in which the reference of the expression is initially fixed. There, an object or a kind must obviously somehow be singled out for naming. According to Kripke, this can happen either with the help of an *ostension* (i.e. by pointing to it or exhibiting it) *or* a *description*. Kripke even adds, '[t]he case of baptism by ostension can perhaps be subsumed under the description concept also. Thus the primary applicability of the description theory is that of initial baptism.' (Kripke 1972/1980, p. 96, footnote 42) The second, very important part of Kripke's picture is the idea of the subsequent transmission of the name, '*reference borrowing*': other language users not present at the name-giving occasion acquire the name and the ability to refer with it from those in attendance at the baptism, still others from the former users, and so on. 'Through various sorts of talk the name is spread from link to link as if by a chain' (Kripke 1972/1980, p. 91). Later users of the expression need not know or be able to identify the referent; it is sufficient, according to Kripke's picture, for successfully referring that they are part of an adequate 'historical' or 'causal' *chain* of language users which goes back to the first users. Speakers may also be largely ignorant of this chain or even forget from whom they got the name.

Even if the expression was originally introduced in the short term by means of a description, the crucial Kripkean thesis is that the description need not retain any long-term association with the expression, and that the expression does not thereby become synonymous with that description. Neither that particular description nor any other uniquely identifying description is generally transmitted – nor need be transmitted – along with the expression. Nevertheless, it appears that these later users can use the expression to refer successfully.[9]

At this point, at least H&W qualify their approach: 'Since reference/extension borrowing is parasitic on reference/extension fixing [baptism], *our primary concern* shall be with the latter' (p. 914, footnote 5; my emphasis). Consequently, H&W say virtually nothing about reference borrowing and later uses of an expression, and focus on the initial introduction, or baptism, of natural kind terms.

3 Further developments of NTR

When the strengths and weaknesses of NTR are evaluated, fairness requires one to take into account also later advances in the theory, and not to focus solely on the earliest, rather sketchy formulations of NTR. In particular, three issues are relevant for our discussion:

[8] For this terminology and classification of arguments, see e.g. Devitt and Sterelny 1987, 1999.
[9] I owe this way of formulating Kripke's key conclusions to Burgess 2013.

First, Field (1973) has argued that some scientific terms may have been, at some point in their history, referentially *indeterminate*, and introduced different notions of *partial reference*. (Field gives, as a possible example, 'mass' in Newtonian physics; according to him, it partially referred to both relativistic mass and proper mass, which are, in the light of more advanced physics, two distinct physical quantities.) Generalising the idea, Devitt (1981) has suggested that the causal-historical theory of reference should be complemented with the idea of partial designation. An expression may in that way, at some period of time, partially refer to more than one different (though partially overlapping) extension.

Second, in order to allow and explain changes of reference, Devitt (1974, 1981) has suggested the idea of '*multiple grounding*' – that it is not only the initial baptism that determines the reference: a name typically becomes multiply grounded in its bearer in other uses of the word relevantly similar to the baptism. In other words, also other uses may involve the application of the word to the object in perceptual confrontation with it (see Devitt 1981, pp. 57–58; Devitt and Sterelny 1999, pp. 75–76). This might result in a shift in extension, but it may also make the reference of an expression more determinate as time passes. (The notion of partial reference may play a role here.) It seems that via a similar process, a vernacular kind term based on superficial properties may also gradually transform into a natural kind term (in some sense).

Third, it has been long recognised among the advocates of NTR that, especially in the case of kind terms, the introduction of a term must involve *some descriptive element* (see e.g. Sterelny 1983; Devitt and Sterelny 1987, 1999; Stanford and Kitcher 2000). Namely, a single sample will usually be simultaneously a member of many kinds. So how can *a general term* such as 'tiger' be introduced? If it happens through an initial baptism in the contact with a sample, as NTR seems to suggest, how can one rule out incorrect kinds of generalisations? This is the so-called *qua* problem. Devitt and Sterelny, for example, granted that some *categorial description* may be used even in the case of proper names, which may in part rule out the wrong sort of generalisations. As Sankey (1994, p. 71) points out, arguably such descriptive elements were already a feature of Putnam's original account of kind terms. Allowing such a descriptive element does not amount to a return to descriptivism. Accordingly, Devitt and Sterelny write:

> Clearly, we have moved some distance back toward the description theories rejected earlier… However, the extent of the move should not be exaggerated. First, the association of a general categorial term certainly does not amount to identifying knowledge of the object. Second, our movement is a modification of the causal theory of grounding [i.e. name introduction]. The causal theory of reference borrowing remains unchanged; borrowers do not have to associate the correct categorical term (Devitt and Sterelny 1987, p. 65).

Stanford and Kitcher (2000) in particular have substantially improved on Putnam's original account of the reference of natural kind terms. Roughly, in their approach, there is a whole range of samples (not only a single sample), a range of foils, and *some associated properties* involved in the introduction of a natural kind term. This shows how one can rule out the wrong kind of generalisation (or at least many of them). According to the approach of Stanford and Kitcher, term introducers make stabs in the dark: they see some observable

properties that are regularly associated, and *conjecture* that *some* underlying property (or 'inner structure') figures as a common constituent of the total causes of each of the properties. This conjecture may be incorrect, in which case the term may fail to refer (or, its role is reconsidered). However, if it is correct, one can exclude incorrect generalisations and fix the reference in the intended way to the set of things that share that underlying property, belong to the same species, etc. The assumption, which sometimes seems to be at work in these debates, that the term must either be completely descriptive or involve no descriptive element at all even in its introduction is a false dichotomy.

Consequently, NTR does *not* necessarily require the extension of a kind term to be absolutely determinate, even across all possible worlds, especially beginning already from the first introduction of the term. The extension may well be somewhat indeterminate, and there may be unclear borderline cases. It is also possible that as science advances, there is some room for negotiation, conventional choice, and stipulation as to which way to go – a possibility that H&W and also LaPorte (2004), for example, underline. Granting such things does not as such mean that NTR has to be false, nor that one has no choice but to accept descriptivism.

4 Varieties of essentialism

A notable share of the criticism of NTR in the philosophy of science is related to micro-essentialism. H&W, for their part, define micro-essentialism thus:

> On micro-essentialism, members or samples of natural kinds are unified by sharing a common micro-structure which (i) explains their macroscopic properties, (ii) holds universally for those members, and (iii) is necessary throughout modal space. (p. 916)

I want to suggest that it is useful to distinguish weaker and stronger versions of essentialism. Namely, straightforward, strong micro-essentialism typically commits itself to the following (cf. Khalidi 2016):

> (SME) An essence of a kind consists of a set of *intrinsic micro-structural* properties which are both *necessary* and *sufficient* for the membership of the kind.

In addition, it is not uncommon to require that all this results in completely sharp boundaries between kinds, totally determinate across the space of all possible worlds. This (SME and sharp boundaries) is the view that H&W and many others attack. Sometimes (e.g. Ellis 2002) it is moreover assumed that natural kinds are (a) hierarchical, (b) discrete, and (c) absolute, i.e. independent from interests.[10] It is true that, as many critics including H&W have pointed out, such strong assumptions arguably fail

[10] In like spirit, Beebee (2013) defines, for her own critical purposes, 'natural kind essentialism' as requiring that kinds are absolute and hierarchical, and that their essences are intrinsic. Although her aims are quite different from mine, she agrees with what I next contend, that we should *not* take it for granted that natural kinds, in this context, necessarily have to be hierarchical and absolute.

in general: SME seems to fail especially for biological kinds; while (a)–(c) arguably fail even for chemical kinds (see e.g. Hendry 2012).

I think H&W and others are right that we should take the talk about such absolute essences of kinds and about a determinate space of possible worlds at least with a grain of salt. Possible worlds are useful heuristic tools, but should perhaps not be assumed, especially in the case of natural kinds, to be well-defined in all sorts of fantastic scenarios where even the laws of nature are very different.[11] We should not, though, throw the baby out with the bathwater. In particular, I contend that none of this shows that we should accept descriptivism after all.

A more minimal and flexible understanding of 'essences' may well require none of the above claims, i.e. SME, sharp boundaries, and (a)–(c): 'underlying features' (what externalism and NTR are commonly leaning on) may be many things, and not necessarily intrinsic micro-essences. Instead, an 'essence' (loosely speaking) may be allowed which depends (at least partly) on *relational* properties, for example, on the right sort of *historical* relations to predecessors. Some biological kinds may be an example (cf. Griffiths 1999; Okasha 2002; Godman 2018). Instead of consisting of straightforward necessary and sufficient conditions universally shared by all the members, the 'essence' (loosely speaking) of a kind may perhaps in some cases have the nature of a *cluster*: different members of the kind may have different properties of the cluster, and no property (or a conjunction of properties) is both sufficient for the membership and necessarily possessed by every member of the kind.[12] As we have seen, it is possible to adopt Field's idea of partial reference, and it is not compulsory to assume that there are absolutely sharp boundaries; *indeterminate* borderline cases can well be allowed. Whether such things really deserve to be called 'essences' is largely a verbal issue and irrelevant here. What matters is that such loosely and minimally understood 'essences', whatever one prefers to call them, are perfectly sufficient for the purposes of NTR and its arguments from ignorance and error.

The critics of NTR do not typically present much in the way of arguments against 'essences' interpreted in such a flexible and minimal way.[13] It is true that many philosophers of science have a dim view of strong, intrinsic micro-essences – especially if applied also to biological kinds. However, it would be an exaggeration to say that the kind of more flexible, minimal notion of 'essence' sketched above would be likewise 'widely rejected among philosophers of science'. Many philosophers of biology, for example, seem to be quite happy with such 'essences'.

[11] This is because the identity of some natural kinds is arguably tied to the actual laws of nature. If those laws are varied, the identity of the kind itself becomes blurry. H&W do not emphasise this point, but Wikforss has pressed it elsewhere (see Wikforss 2013), and it seems to be for its part behind what H&W say about modal space. The idea goes back at least to (later) Putnam (1990). I am inclined to accept the point.

[12] Does this amount to the cluster theory version of descriptivism? It does not, because according to the view proposed here, it may well be that no language user nor even the linguistic community as a whole knows, at a given time, the cluster, or even substantial parts of it; the correct cluster may be instead discovered a posteriori step by step in the future as empirical science advances. Cf. Section 7.

[13] H&W do mention a particular well-developed variant of cluster essences, Boyd's Homeostatic Property Cluster Theory (HPCT); they note that it has also been criticised in the literature, and that it does not generalise. However, H&W seem themselves to sympathise with some kind of cluster idea. So the pressing questions for them are: What, more exactly, is the relevant cluster supposed to be like? Is it assumed that a kind term becomes synonymous with the cluster of descriptions used already in the introduction of the term? Is that cluster (or substantial parts of it) generally associated with the term by language users? Cf. Section 5.

In general, NTR need not necessarily assume, as many opponents tacitly presuppose in their critique, that all kind terms in various different areas of knowledge, such as chemistry, biology, and fundamental physics, should function in exactly the same way: the success of NTR in no way requires an overarching, universal, and unified micro-essentialist theory of natural kinds.[14] What are essential for NTR are plausible counterexamples for descriptivism (or internalism) and the phenomenon of ignorance and error in particular. It is quite consistent with NTR that what exactly counts as a natural kind and how their 'essence' (loosely speaking) is determined varies across different areas of knowledge. For example, systematising certain scattered remarks by Putnam, I have suggested (see Raatikainen 2020) that there are at least four different sorts of referring expressions, which mean and refer in unlike ways. I think that there are in particular scientific kind terms of all four of those categories. Thus, the picture NTR suggests applies best to kinds which are first identified in perceptual confrontation with samples, but it is allowed that some underlying features may play a role in determining its extension.[15] Extremely theoretical kind terms of fundamental physics, for example, where there is no such perceptual contact, may in contrast well be descriptive. Then again, some kind terms may also be entirely observational. There is no reason to try to cast all such quite diverse kinds in the same mould.

Moreover, it is worth noticing that 'natural kind' is not unambiguous: sometimes it is used very broadly for all sorts of kinds in natural sciences, only to contrast them with artefact kinds (such as *chair*), arbitrary kinds (e.g. *less than 1.32 m high*), or kinds that are gerrymandered and artificial (like Goodman's *grue*); observational kinds such as *yellow* may then count as natural kinds. There is no reason to assume that they all would have that much philosophically interesting in common. At times the focus is on kinds that support inductive inference; their scope is likewise wide. Occasionally, a much more specific notion is at stake; the concept may be highly theory-laden, and different philosophical accounts have their own notions of natural kind. Confusions easily result if this variety of uses is not recognised.

As an alternative to the strong and naïve metaphysical picture that many critics including H&W attribute to the advocates of NTR, one might adopt, for example, the pluralistic 'promiscuous realism' put forward by Dupré (1993). According to it, there are countless ways of taxonomising the world into kinds; the structure of the world is vastly complex and can be categorised in a number of different cross-cutting ways according to the different scientific and other interests we happen to be pursuing. Nevertheless, this is not a conventionalist view, but realism about natural kinds. A kind may be relative to an interest, but once the interest is fixed, the kind is quite objective; it may have vague boundaries, but its demarcation need not reduce to readily observable properties, and ignorance and error are perfectly possible. This is all that NTR in reality needs: naïve and strong essentialism is in no way required.

[14] Obviously, I am not denying that some enthusiasts of NTR seem to have in practice assumed that 'natural kind' is a more unified concept than it actually is. Here I am only talking about what NTR, arguably, necessarily must assume and what it need not assume.

[15] These are sometimes called 'manifest kinds' or 'observational natural kinds' in the literature.

5 The resurrection of descriptivism?

As to the juxtaposition of NTR and descriptivism, the following three questions are repeatedly conflated in the critical literature, and should be more clearly distinguished:

(1) Can a description (or a cluster of descriptions) be used in the introduction ('baptism') of a term to initially fix its reference/extension?
(2) If so, does the term thereby become synonymous with that description?
(3) Is that description passed from speaker to speaker such that any 'competent' language user would associate that description with the term?

H&W and many other opponents of NTR argue primarily for the positive answer to (1) (H&W explicitly qualify their argument that way; see Section 2 above). However, (1) is perfectly compatible with everything that Kripke, Putnam, and others say: as we have seen, descriptions can well be used, in the picture that NTR proposes, to single out an entity or a kind in the initial baptism. Granting this utterly does not amount to descriptivism. Kripke, Putnam, and others rather focus on arguing against (2) and (3). Many critics of NTR, including H&W, by contrast, say practically nothing about them. Nevertheless, affirmative answers to these latter questions are critical for the survival of descriptivism.

Externalism in particular is, as the title of Putnam's seminal paper 'Meaning of "meaning"' (Putnam 1975) already makes apparent, an account of *meaning*, and not of the mere introduction of an expression. The standard versions of descriptivism, which aim to solve Frege's puzzles facing the Millian theory of *meaning*, are likewise theories of *meaning*, and of what an average language-user knows. H&W also grant this, as they define, in the very beginning, descriptivism as the view that 'the *meanings* of kind terms are given by a set of associated descriptions' (pp. 911–12; my emphasis). But in order to defend *that* view, it is just not enough to argue that descriptions are used in the introduction of kind terms; hardly anyone denies that. It should be demonstrated that the term thereby becomes *synonymous* with that description, and that the later average language users regularly associate *that* description with the word. H&W, for example, do not present anything in defence of these latter claims, which are essential for descriptivism, and the critical arguments of Kripke, Putnam, and others cast doubt on them.

6 A more sophisticated descriptivism?

Wikforss has contended time and again that in the case of kind terms, we should adopt descriptivism and more exactly a cluster theory of a more sophisticated sort (see Wikforss 2005, 2008, 2013). H&W restate the claim and summarise the idea as follows:

> *A sophisticated cluster theory* would depart from standard versions of the theory in at least two respects: First, legitimate descriptions are *not* restricted to *observable properties* – the cluster theory in itself is not committed to 'superficialism'

> [footnote: See (Lewis [1994], p. 424]. Second, the theory would not commit to the idea that all descriptions are given equal weight. (p. 928; my emphasis)

However, the idea of weighting the descriptions was already critically reflected on by Kripke, and plausibly argued not to make much difference in the face of the critical arguments. Be that as it may, Wikforss has constantly argued for a version of descriptivism in which the relevant descriptions would not be restricted to superficial properties (see Wikforss 2005, 2008, 2013), and H&W reiterate the proposal. For further clarifications, they defer to Lewis. This is how he, in turn, explains it (the example term is, again, 'water'):

> ...there is more to the cluster than that [it is liquid, it is colourless, it is odorless, it supports life]. Another condition is: it is a *natural kind*. Another condition is *indexical*: it is abundant hereabouts. Another is *metalinguistic*: many call it 'water'. (Lewis 1994, p. 313; my emphasis)

But how much could such an expansion of descriptions in the cluster actually aid descriptivists such as H&W? Not much, it seems. To begin with, it is not clear that H&W can really help themselves with Lewis' first addition, i.e. 'is a natural kind', for H&W contend that there are no such things as natural kinds, as standardly understood. It seems that Lewis, even if he wanted to defend sophisticated descriptivism, more or less took for granted the neo-classical understanding of natural kinds that H&W aim to rebut.[16] In any case, this condition could rule out, at best, some artificial and gerrymandered kinds, and would be of no help in discriminating one natural kind from another, superficially indistinguishable natural kind (such as Putnam's imagined twin-water XYZ).

Moreover, H&W cannot lean on metalinguistic descriptions either. First, recall that H&W state that their primary concern is the initial *introduction* of a new kind term, its 'baptism'. But at the time, the kind is not (typically) yet called anything, and hence its being called (whatever) '*K*' cannot be utilised in the cluster of descriptions to single out the kind for naming. Second, even if one was rather interested in later uses of the term, arguably metalinguistic descriptions are just completely the wrong sort of descriptions for the job of expressing *meanings* (see e.g. Raatikainen 2020). H&W, among others, grant the standard view that descriptivism is a theory of meaning. Third, *if* the aim is rather to defend internalism, not just any description will do: they must be such that the mental state resulting from associating the descriptions with the expression is a narrow state. Metalinguistic descriptions apparently violate this condition.[17]

Finally, indexicals may be undoubtably used, for their part, in the introduction of a kind word, to single out a kind for naming, to initially fix its extension (at least roughly). Nothing that Kripke and Putnam (and others) assert rules this out, though. Indeed, they explicitly allow this. The further and crucial question, however, is whether the kind term thereby becomes permanently synonymous with that indexical-involving

[16] H&W write that in the sort of semantic theory they favour, 'there will no longer be a distinct semantic category of natural kind terms', and that this 'harmonizes with the rejection of a sharp metaphysical demarcation between natural kinds and other kinds' (pp. 928–29); cf. Wikforss 2010.

[17] It would take me too far afield to argue the matter in detail. Moreover, this point is not particularly important for my overall argument. Therefore, I must leave the more thorough discussion of this issue for another occasion.

description, and hence itself essentially indexical. There are many good reasons to think that our common kind words are not themselves indexical. Indexicals can be utilised in the beginning, in focusing on the kind, without making the resulting expression itself indexical (just like one can use a description involving an indexical, say, 'this baby', in the introduction of a proper name; this does not make the name itself indexical[18]; cf. the difference between (1) and (2)–(3) in the Section 5 above).[19]

The *meaning* of a kind term should not in the first place be tied too tightly to the contingent circumstances of the term's introduction. For example, we can imagine a future scenario in which the Earth and all its water has been destroyed in a cosmic catastrophe, and the remaining people live in a space station and get their scant water transported from asteroids in the form of ice. However, it is not clear that the *meaning* of 'water' should thereby have changed. And if so, the *meaning* of 'water' cannot be anything that descriptions such as 'The clear liquid that fills seas, lakes, and rivers, falls from the sky in rain… (etc.) abundant *here*' express. For comparison, consider, for example, *uranium*: it was first identified as something that gives a specific yellow colour to glass; that was its only use for centuries, beginning from Ancient Rome. Few people today think about that property when they use the word 'uranium'. It is not plausible that the capability to give a specific yellow colour to glass would somehow analytically belong to the *meaning* of 'uranium'. In 1789, Klaproth officially discovered uranium and named it, but he had no idea of its radioactivity, the properties of which had not yet been discovered and would occur only a century later. Today we routinely associate radiation with 'uranium', but radioactivity has not been analytically associated with it from the beginning.

In sum, the 'sophisticated' version of descriptivism that would go beyond 'superficialism', which H&W suggest we should adopt, remains dissatisfyingly unclear. It is quite difficult to see what kind of version would serve their philosophical purposes and also be plausible as a theory of shared meaning. At the very least, the advocates of descriptivism have the burden of providing a much more detailed account of what exactly such a new sophisticated version of descriptivism would look like – and what philosophical work precisely it is supposed to do.

7 What of the common concept strategy?

It is instructive to reflect on what has been a very popular countermove against NTR, namely what is sometimes called 'the common concept strategy'. A repeated response to Putnam's Twin Earth thought experiment has been the following suggestion: why not simply say that our word 'water' has (in the scenario in which Twin Earth exists and XYZ plays the 'water role' there) – at least when nobody knows the chemical constitution of water – in its extension both H_2O and the qualitatively similar XYZ?[20]

[18] Let us imagine that most human names were introduced with this description, i.e. 'this baby'. If all those millions of names would become synonymous with this description, they would all have the same meaning as each other – even if they all denoted distinct persons. This seems just preposterous.

[19] Wikforss herself discusses the indexical response more elsewhere, and not entirely uncritically; see Wikforss 2008, p. 169.

[20] See e.g. Zemach 1976; Mellor 1977; Searle 1983, p. 203; Bach 1987, p. 276; Segal 2000, p. 19. Wikforss (2008) too reflects on the response more or less sympathetically but does not univocally commit herself to it.

There is, however, a quite convincing response to this strategy, due to McCulloch (1992, 2003): one may flirt with this reply strategy in a particular case (e.g., with 'water' and XYZ), but all that NTR really needs is the possibility of a case where everything appears the same but nevertheless two speakers talk about different things. In the thought experiment, XYZ can be *stipulated* to be a substance *radically different* from water and only superficially similar to it, and hence definitely a different substance – it is *stipulated* that it is *not* water. Or, if the particular example of water seems unclear, one can switch it to, say, gold – and some other (hypothetical) yellow, shiny, and malleable metal-like stuff, which is stipulated *not* to be gold.[21] Such a radically different kind should not be conflated with mere borderline cases.

As McCulloch points out, denying generally the very possibility of such a case apparently amounts to the claim that there can be no difference in kinds without some readily observable difference – that everything that looks like the same kind really is the same kind. However, the latter is a strong and controversial empiricist thesis which is, at best, in need of a substantial defence and certainly cannot be taken for granted (see McCulloch 1992, 2003). (I would add that 'observable' here should understood as *observable by the methods of some fixed time*, e.g. 1750.) It is hardly true that such a view is widely held by philosophers of science today. Nothing here requires strong, naïve, and overarching essentialism and absolute, intrinsic micro-essences with sharp boundaries. The difference may depend, for example, on historical origin or other hidden relational properties, and the boundaries may well be somewhat indeterminate.

Even if a kind does not have a straightforward micro-essence – that is, not all its members or samples share exactly the same intrinsic micro-structure or such – a large enough deviance in relevant non-observable properties may nevertheless count as a sufficient reason for disregarding something as belonging to the kind. Such things do actually happen in science. For example, the African elephant was long considered a single species. However, recent detailed genetic studies (around 2010) have led scientists to conclude that there have been undeniably two separate species under that single label all along – the African bush elephant and the African forest elephant – which are just as distinct from each other as the Asian elephant and the woolly mammoth. Such a discovery would not even have been possible with the methods available some decades earlier.

It is not clear where exactly H&W, for example, stand with respect to such questions. On the one hand, they distance themselves from 'superficialism', suggesting that they would, more or less, agree here. On the other hand, they state that they are focusing on the introduction of a kind term. However, the relevant non-superficial properties are typically discovered only in the later stages of research and are not known at the time of the initial introduction of the term: they are not typically analytically associated with the term from the beginning. It may well be that the 'essence' (loosely speaking) of some kind really is truthfully described as a cluster of properties rather than as a straightforward micro-essence. However, it is plausible that such clusters are revealed only step by step, as scientific research advances, and are not

[21] As H&W and some earlier critics of NTR point out in some detail, water turns out to be, chemistry-wise, a rather complicated case (but see Hendry 2006; Hoefer and Martí 2019). However, gold, by contrast, behaves quite nicely.

available at the moment of the first introduction of the relevant kind term. This makes the views, such as those of H&W, quite unstable.

I contend that the questions truly crucial in relation to NTR are the following:

(A) Could there be a difference of a kind, and a difference in the extension of a kind term, which is not observable and not yet detectible by the methods available at the given time?
(B) Is it possible to introduce a kind term in a specific context, and is it possible to refer with it successfully (within some tolerable degree of indeterminacy and partial reference), even if there is not yet a reliable method available for verifying generically whether a given sample belongs to the kind or not?

Unless a strong verificationist variant of empiricism is presupposed, a positive answer to both questions seems plausible. But this is all that NTR and externalism need. (A) and (B) do not require strong intrinsic micro-essences. They do not involve any strong assumptions about distant, merely metaphysically possible worlds and how the extension of a term behaves there. Even modal arguments against descriptivism are principally negative: they aim to undermine the thesis that a name is synonymous with a (cluster of) description. They do not require a well-defined space of all metaphysically possible worlds and in particular determinate extensions in distant worlds, but only a plausible counterexample. The arguments from ignorance and error, most central to NTR, presuppose even less.

8 Some historical examples

Lastly, let us reflect briefly on some historical examples, which apparently harmonise with the picture sketched above and suggest that it is more than just abstract philosophical speculation.

Gold The first known gold artefacts are from Ancient Egypt, from the 5th–4th millennium BC. For the sake of argument, let us assume that our concept of gold derives from Egypt, where the symbol 'Nebu' was used to denote it. In addition to its obvious properties – being yellow, shiny, dense, and malleable – the Egyptians associated gold with the sun god Ra, and believed that gold is a *heavenly* and *indestructible* substance. However, it would be rather absurd to suggest that the latter properties belong analytically to the meaning of *our* word 'gold', even if they perhaps were firmly associated with the predecessor of our word 'gold' in the early uses of that word.

'Mountain sickness' From the sixteenth century onwards, miners in certain specific regions in Germany and Czechoslovakia regularly died of a mysterious disease which was soon called 'mountain sickness'. It was earlier assumed that the ores mined were responsible, arsenic in particular. In 1879, the disease was identified as lung cancer, and the received view now is that it was caused by the high levels of radon in the mines. However, radon was discovered only in 1899, and it was properly isolated and its many key properties determined only in 1909. Radioactivity in general was discovered in 1896. The awareness of the connection between cancer and ionising radiation emerged

gradually only few decades later. The nature, the 'essence', so to say, of mountain sickness was discovered, step by step, many centuries after the initial identification of the disease.

'The silent killer': Carbon monoxide The earliest descriptions of (what clearly were) carbon monoxide poisonings in the vicinity of fireplaces go back to Ancient Greece and Rome (however, the first theories of what was happening were rather wild). The thirteenth-century alchemist Arnold of Villanova described an invisible poisonous gas produced by the incomplete combustion of wood and was undoubtedly talking about carbon monoxide. Carbon monoxide was first prepared artificially by De Lassone in 1776 by heating zinc oxide with coke. However, as it burned with a blue flame, he mistakenly took the gas he had produced to be hydrogen (which was already known from other contexts). The discovery of carbon monoxide is often credited to Joseph Priestley. In the last three decades of the eighteenth century, Priestley gradually recognised the specific properties of this compound and how it was different from (what we know as) carbon dioxide, with which it often appeared. However, Priestley was a firm believer in the phlogiston theory, so his overall analysis could not be quite accurate. In 1800, Cruikshank showed that it was a compound containing carbon and oxygen. Clearly, people were able to refer to this specific poisonous gas some time before its true nature was discovered, and it plausibly had a sort of (loosely speaking) 'essence'.

Oxygen (O) was properly discovered independently by Scheele and Priestley in the 1770s. Sheele called it 'fire air'; Priestley 'dephlogisticated air'. Lavoisier soon after rebutted the whole phlogiston theory of combustion. Shortly after he renamed what he had first called 'vital air' 'oxygen', from the Greek; *oxys*, 'acid', and -genēs, 'producer'. This is because Lavoisier was convinced that oxygen is a common constituent of all acids. This was eventually shown to be wrong: it was rather hydrogen that is common to all acids. Nevertheless, Sheele, Priestley, and Lavoisier were seemingly able to refer successfully to oxygen, even if some of the central beliefs they associated with it were false.[22] Around 1783, Cavendish and Watts independently concluded that hydrogen and oxygen are the constituents of *water*. This must have been a surprise for those who were already familiar with oxygen (and, obviously, with water).

Potash and soda In his 1789 textbook *Traité*, Lavoisier omitted potash and soda from the list of simple substances and wrote that they are 'evidently composed, although we are still ignorant of the principles that enter their composition'. In 1807, Davy confirmed that Lavoisier had been right: both contained oxygen combined with potassium and sodium (recently discovered elements which Davy isolated at the same time via electrolysis, a newly developed technique). Lavoisier was apparently able to refer to potash and soda and hypothesise that they have some inner nature, even if he did not at the time possess a method for determining what it was.

[22] The case of oxygen is discussed in more detail in Hendry 2010.

All of these examples cohere well with the general picture suggested above that kinds can be and have been named in a specific context, and people have been able to successfully refer to these kinds before their 'nature' had been discovered, in the absence of accurate descriptions, sometimes even with some central false beliefs about the nature. Early users of a kind name, even experts, may well misclassify some samples. One may lack, at the time, the concrete means for deciding the issue reliably. There are all sorts of cases of ignorance and error. That is, these examples speak, for their part, in favour of a positive answer to (A) and (B) in the previous section. I contend that it is quite difficult make clear sense of such episodes if one does not grant the basic ideas of NTR; these would all have been just repeated changes of meaning, and no cumulation of knowledge of the relevant kinds.

9 Conclusions

Many opponents, including H&W saddle NTR with certain strong views on essences, which makes the resulting view an easy target for criticism. It is true that few philosophers of science would endorse such a combination of views. Although a critical discussion of naïve, straightforward, and overarching microessentialism is certainly in order, it is a mistake to conclude that NTR in general would therefore be doomed, and that there is no choice but to return to descriptivism. Arguably NTR does not necessarily require such a strong and overarching version of essentialism, and descriptivism remains at any rate problematic. Any even modestly realistic picture of kinds – be it however pluralistic, for example – supports arguments from ignorance and error, and that is quite enough for the plausibility of NTR.

Acknowledgements I have been engaged in a friendly argument on these themes with Sören Häggqvist and Åsa Wikforss for more than a decade. Although we still seem to disagree on some issues, I have benefited enormously from this discussion: I have revised my views more than once, and in general, my own thoughts have become much clearer in the process. I am grateful to Sören and Åsa for both their contributions to this exchange and the friendly atmosphere of the debate. They presented an earlier version of the paper at stake in Tampere (where I am affiliated) already a few years ago. I did not quite manage to formulate my contrasting view so clearly back then, so this paper could be viewed as my delayed response.

An earlier version of this paper was presented at *The Swedish Congress of Philosophy 2019* in Umeå, Sweden, in June 2019. I would like to thank the organisers for the invitation, and all of those in attendance who participated in the discussion, especially Sören for his detailed comments.

I have also benefited from discussions and correspondence with Genoveva Martí, Carl Hoefer, Robin Hendry, and Michael Devitt on these topics, and I would like to express my gratitude to them.

I would also like to thank the anonymous referees for their valuable comments and suggestions that definitely improved this paper.

Compliance with ethical standards

Conflict of interest The author declares that the author has no conflicts of interest related to the research for this paper or its contents.

Ethical approval not applicable.

Informed consent not applicable.

References

Bach, K. (1987). *Thought and reference*. Oxford: Oxford University Press.

Beebee, H. (2013). How to carve across the joints in nature without abandoning Kripke–Putnam semantics. In S. Mumford & M. Tugby (Eds.), *Metaphysics and science* (pp. 141–163). Oxford: Oxford University Press.

Braun, D. (2006). Names and natural kind terms. In E. Lepore & B. C. Smith (Eds.), *Oxford handbook of philosophy of language* (pp. 490–515). Oxford: Oxford University Press.

Burge, T. (1979). Individualism and the mental. In P. French, T. Uehling, & H. Wettstein (Eds.), *Midwest studies in philosophy, 4* (pp. 73–121). Minneapolis: University of Minnesota Press.

Burge, T. (1982). Other bodies. In A. Woodfield (Ed.), *Thought and object: Essays on intentionality* (pp. 97–120). Oxford: Oxford University Press.

Burgess, J. P. (2013). *Kripke. Puzzles and mysteries*. Key contemporary thinkers. Cambridge: Polity.

Devitt, M. (1974). Singular terms. *Journal of Philosophy, 71*, 183–205.

Devitt, M. (1981). *Designation*. New York: Columbia University Press.

Devitt, M. (1989). Against direct reference. *Midwest Studies in Philosophy, 14*, 206–240.

Devitt, M. (1990). Meanings just ain't in the head. In G. Boolos (Ed.), *Meaning and method: Essays in honor of Hilary Putnam* (pp. 79–104). Cambridge: Cambridge University Press.

Devitt, M., & Sterelny, K. (1987). *Language and reality*. Oxford: Basil Blackwell.

Devitt, M., & Sterelny, K. (1999). *Language and reality* (Second ed.). Oxford: Blackwell.

Donnellan, K. (1970). Proper names and identifying descriptions. *Synthese,* 21, 335–358. Reprinted in D. Davidson & G. Harman (Eds.), *Semantics of natural language* (pp. 356–79). Dordrecht: Reidel, 1972.

Dupré, J. (1981). Natural kinds and biological taxa. *Philosophical Review, 90*, 66–90.

Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.

Ellis, B. (2002). *The philosophy of nature: A guide to the new essentialism*. Chesham: Acumen.

Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy, 70*, 462–481.

Godman, M. (2018). Scientific realism with historical essences: The case of species. *Synthese*. https://doi.org/10.1007/s11229-018-02034-3.

Griffiths, P. E. (1999). Squaring the circle: Natural kinds with historical essences. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 209–228). Cambridge, MA: MIT Press.

Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies, 61*, 109–126.

Hacking, I. (2007). Natural kinds: Rosy dawn, scholastic twilight. *Royal Institute of Philosophy Supplement, 61*, 203–239.

Häggqvist, S., & Wikforss, Å. (2018). Natural kinds and natural kind terms: Myth and reality. *British Journal for the Philosophy of Science, 69*, 911–933.

Hendry, R. F. (2006). Elements, compounds and other chemical kinds. *Philosophy of Science, 73*, 864–875.

Hendry, R. F. (2010). The elements and conceptual change. In H. Beebee & N. Sabbarton-Leary (Eds.), *The semantics and metaphysics of natural kinds* (pp. 137–158). New York: Routledge.

Hendry, R. F. (2012). Chemical substances and the limits of pluralism. *Foundations of Chemistry, 14*, 55–68.

Hoefer, C., & Martí, G. (2019). Water has a microstructural essence after all. *European Journal for Philosophy of Science, 9*(1), 12.

Hoefer, C., & Martí, G. (2020). Realism, reference & perspective. *European Journal for Philosophy of Science, 10*(3), 1–22.

Khalidi, M. A. (2016). Natural kinds. In P. Humphreys (Ed.), *Oxford handbook of the philosophy of science* (pp. 397–416). Oxford: Oxford University Press.

Kripke, S. (1972/1980). Naming and necessity. In D. Davidson & G. Harman (Eds.), *Semantics of natural language* (pp. 253–355). Dordrecht: Reidel, 1972. Reprinted with a new introduction: Saul Kripke, *Naming and necessity*. Cambridge, MA: Harvard University Press, 1980.

LaPorte, J. (1996). Chemical kind term reference and the discovery of essence. *Noûs, 30*, 112–132.

LaPorte, J. (2004). *Natural kinds and conceptual change*. Cambridge: Cambridge University Press.

Lewis, D. (1994). Lewis, David. In S. Guttenplan (Ed.), *A companion to philosophy of mind* (pp. 412–431). Oxford: Blackwell publishers. Reprinted, with the title 'reduction of mind', in D. Lewis, *Papers in metaphysics and epistemology* (pp. 291–324). Cambridge: Cambridge University Press, 1999. (page references are to the reprint.)

McCulloch, G. (1992). The spirit of twin earth. *Analysis, 52*, 168–174.

McCulloch, G. (2003). *The life of the mind*. London: Routledge.

Mellor, D. H. (1977). Natural kinds. *The British Journal for the Philosophy of Science, 28*, 299–312.

Needham, P. (2000). What is water? *Analysis, 60*, 13–21.

Needham, P. (2002). The discovery that water is H_2O. *International Studies in the Philosophy of Science, 16*, 205–226.

Needham, P. (2011). Microessentialism: What is the argument? *Noûs, 45*, 1–21.

Okasha, S. (2002). Darwinian metaphysics: Species and the question of essentialism. *Synthese, 131*, 191–213.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (Ed.), *Language, mind, and knowledge. Minnesota studies in the philosophy of science VII* (pp. 131–193). Minneapolis: University of Minnesota Press.

Putnam, H. (1990). Is water necessarily H_2O? In H. Putnam (Ed.), *Realism with a human face* (pp. 54–79). Cambridge: Harvard University Press.

Raatikainen, P. (2020). Theories of reference: What was the question? In A. Bianchi (Ed.), *Language and reality from a naturalistic perspective: Themes from Michael Devitt* (pp. 69–103). Cham: Springer.

Sankey, H. (1994). *The incommensurability thesis*. Aldershot: Avebury.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Segal, G. (2000). *A slim book about narrow content*. Cambridge: MIT Press.

Stanford, P. K., & Kitcher, P. (2000). Refining the causal theory of reference for natural kind terms. *Philosophical Studies, 97*, 97–127.

Sterelny, K. (1983). Natural kind terms. *Pacific Philosophical Quarterly, 64*, 110–125.

Wikforss, Å. (2005). Naming natural kinds. *Synthese, 145*, 65–87.

Wikforss, Å. (2008). Semantic externalism and psychological externalism. *Philosophy Compass, 3*, 158–181.

Wikforss, Å. (2010). Are natural kind terms special? In H. Beebee & N. Sabbarton-Leary (Eds.), *The semantics and metaphysics of natural kinds* (pp. 64–83). New York: Routledge.

Wikforss, Å. (2013). Bachelors, energy, cats and water: Putnam on kinds and kind terms. *Theoria, 79*, 242–261.

Williamson, T. (2006). Can cognition be factorised into internal and external components? In R. Stainton (Ed.), *Contemporary debates in cognitive science* (pp. 291–306). Oxford: Blackwell.

Zemach, E. (1976). Putnam's theory of reference of substance terms. *Journal of Philosophy, 73*, 116–127.

Yli-Vakkuri, J., & Hawthorne, J. (2018). *Narrow content*. Oxford: Oxford University Press.