Synthese (2024) 203:131
https://doi.org/10.1007/s11229-024-04568-1

ORIGINAL RESEARCH

Defending the pure causal-historical theory of reference fixing for natural kind terms

Jaakko Tapio Reinikainen[1]

Received: 24 August 2023 / Accepted: 15 March 2024

Abstract
According to the causal-historical theory of reference, natural kind terms refer in virtue of complicated causal relations the speakers have to their environment. A common objection to the theory is that purely causal relations are insufficient to fix reference in a determinate fashion. The so-called hybrid view holds that what is also needed for successful fixing are true descriptions associated in the mind of the speaker with the referent. The main claim of this paper is that the objection fails: reference fixing of natural kind terms can be purely causal. The main argument draws inspiration from recent theoretical advances made in metaphysics of kinds by Marion Godman, Antonella Mallozzi, and David Papineau. The main claim is that their notion of super-explanatory properties may explain how reference of many kind terms can be fixed purely causally.

Keywords Reference · Causal-historical theory of reference · Qua-problem · Reference fixing · Super-explanatory properties

1 Introduction

The main aim of this paper is to defend the so-called 'pure' causal-historical answer to the infamous *qua*-problems regarding the reference fixing of natural kind terms. The main opposition here are those views according to which some descriptive content that is true of the kind named by a term *must* be present in reference fixing for determinate reference to succeed. These views include not only classical descriptivism, but also various hybrid views explicated below.

The paper proceeds as follows. Section 2 will clarify the *qua*-problems and the basic hybrid view. The main argument in Sect. 3 presents a coherent just-so story

✉ Jaakko Tapio Reinikainen
jaakko.reinikainen@tuni.fi

[1] Faculty of Social Sciences, Philosophy, Tampere University, Tampere, Finland

Published online: 18 April 2024

how the reference of some natural kind terms can be fixed purely causally, i.e., without mediation of any true description in the mind of the speaker associated with the naming term about the kind or object being named. Section 4 includes a separate argument for how the pure causal-historical account can explain reference failures.

2 The *qua*-problems and the hybrid solution

To begin with, the causal-historical theory of reference that originates from Kripke (1980) and as developed further by Devitt (1996) should be understood as consisting of two basic parts:

Grounding. Following Kripke's lead, a referring term (paradigmatically a proper name) usually becomes fixed in its referent by an 'initial baptism' where the speaker is in perceptual contact with the object and intentionally uses the term to refer to the object. The first such 'grounding' act may be followed by several others in what Devitt calls 'multiple grounding' of the term. (Devitt, 1996, 167)

Borrowing. Once the term has been grounded in a referent, it can be 'borrowed' by speakers who have not been in perceptual contact with the referent from someone who has been, and then borrowed further from those people. While borrowing also is an intentional act, for most terms it is not necessary that the borrower receives a true description of the referent along with the term, or later remembers from whom she received the term. (Devitt, 1996, 164)

In what follows, I will focus exclusively on the grounding part of the theory, which is supposed to answer how a term, most importantly a proper name or a natural kind term, has its reference fixed. In a nutshell, a pure causal-historical account of reference fixing holds that a causal grounding is sufficient to fix the reference of at least some natural kind terms. A causal relation need not be necessary; reference can also be fixed purely descriptively. In contrast, the basic hybrid view of reference fixing holds that a causal grounding relation must *always* be accompanied by a true, non-trivial description of the kind named for the fixing to succeed.[1]

Next, I will explain (i) what the main motivations for hybrid accounts are and (ii) elaborate on the two conditions of *true, non-trivial descriptions* that make up the basic hybrid view.

[1] There is some significant variation in the details of competing hybrid theories that are not relevant for my discussion, for my argument targets the basic hybrid view defined by the two conditions discussed below which all hybrid theories start with. For some prominent hybrid theories of reference fixing, see Cummiskey (1992), Devitt and Sterelny (1999), Stanford and Kitcher (2000), Mallon (2017), Gómez-Torrente (2019). Cummiskey's hybrid theory primarily concerns theoretical terms, which is in line with my own views presented in Sect. 4, but otherwise these hybrid theorists accept the basic hybrid view.

There are two main motivations for hybrid accounts: the *qua*-problems and explaining reference failure.[2] A classic formulation of the *qua*-problem comes from Devitt and Sterelny:

> The term is applied to the sample […] *qua* member of one particular natural kind. Any sample of a natural kind is likely to be a sample of many natural kinds; for example, the sample is not only an echidna, but also a monotreme, a mammal, a vertebrate, and so on. In virtue of what is the grounding in it *qua* member of one natural kind and not another? As a result of groundings, a term refers to all objects having the same underlying nature as the objects in the sample. But which underlying nature? The sample shares many. (1999, 92)[3]

The same problem can be presented to proper names. Why does 'Everest' refer to the mountain, or 'Paris' to the city, even though both Everest and Paris also belong to many other kinds and classes? In principle, all referential terms of a language face some form of the *qua* problem, at least in the view of the causal-historical account. Moreover, the grounding relation of a general term to its referent is never direct as it is with proper names, but is mediated by tokens of the kind, and further by the properties of the tokens. So, why is it that, e.g., 'tiger' is not grounded in some properties of tigers, but in tigers as a kind? Finally, a speaker initiating a grounding is often in causal contact not with a whole object, but only a spatio-temporal slice of it. For example, for a long time, humanity could only observe one side of the Moon; so why is it that what we named (then) was the whole Moon and not one side of it?

As Devitt and Sterelny note, these *qua*-problems arise not because of a serious doubt that our naming practices pick up radically different referents than we ordinarily think, but simply because it is perfectly possible to name a whole object, or one of the several general kinds it belongs to, or one of its properties, or a spatio-temporal slice of it (1999, 79). So, all things being equal, there should be an explanation for why one kind of naming takes place in one case and another in another.

The basic hybrid account provides a simple explanation: the accompanying description, which can be explicit or implicit, is what determines what kind of object is named and excludes other possible references. For example, a speaker who perceives a tiger and says 'I name that *animal species* "tiger"' thereby specifies that what she is naming is the whole animal (species), not another taxon or a spatio-temporal part or a property of the tiger. In this context, it does not matter how exactly the description is 'associated' with the grounding act, e.g. is it consciously thought of or somehow unconsciously present (Devitt & Sterelny, 1999, 80).

Here, we can see why the description associated with the grounding term should be informationally non-trivial. Information is non-trivial when it rules out some other description, and the more descriptions it rules out, the less trivial it is. 'Let that *thing*

[2] I prefer to talk about *qua*-problems, as in plural, because it should not be assumed that all the alternative reference classes can be ruled out by one solution, even though that is what the basic hybrid account in effect claims to achieve.

[3] In his later writings Devitt (2015, 115, fn.) has said that he does not regard the *qua*-problem(s) as at bottom philosophical at all, though they may pose issues to psychology.

be called a "tiger"' is a trivial description because 'thing' does not rule out any other description, hence not any potential reference class. 'Let that *material or physical thing* be called a "tiger"' is on the borderline of triviality: what are the 'non-material, non-physical' things supposed to be ruled out here? In any case, such a description does not rule out the naming of the tiger's stripes, or shape, for instance. So, the associated description must be informative enough to solve the *qua*-problems.

Notice that even a functional description can and should meet these two requirements. In Stanford and Kitcher's hybrid proposal, the associated description includes both a description of the properties of the named samples and a

> *conjecture* that there's some underlying property (or "inner structure") that figures as a common constituent of the total causes of each of the properties. That conjecture could be wrong. But, if it's right, they can beat off Locke's challenge and fix the reference of the term to the set of things that share that underlying property. (Stanford & Kitcher, 2000, 114)

Another core motivation for the basic hybrid account comes from explaining reference failure. It is commonly held that sometimes attempts to fix the reference of a general or singular term fail. This is especially true with theoretical terms like 'phlogiston' and 'Vulcan', but arguably also true with non-theoretical terms like 'witch'. (I won't here address the question how exactly theoretical and non-theoretical terms should be distinguished.) This is a problem for the pure causal-historical theory of reference fixing because even with the grounding of a term like 'phlogiston' or 'witch', the speakers were in causal contact with something observable that prompted their grounding acts. So, it seems that there should be something which these terms refer to after all, according to the theory, although we usually take them to be empty.

The basic hybrid account again has a ready solution available: reference grounding fails when the associated description is false of the kind or object being named (Devitt & Sterelny, 1999, 80). This naturally explains why the associated description must be true of the kind or object named for the fixing to succeed.

While apparently providing simple solutions to two difficult problems, the basic hybrid account comes to conflict with some of Kripke's original thoughts on the causal-historical account. For example, commenting on similar thoughts by Putnam, Kripke said:

> The example he [Putnam] gives is 'cats are animals'. Cats might turn out to be automata, or strange demons (not his example) planted by a magician. Suppose they turned out to be a species of demons. Then on his view, and I think also my view, the inclination is to say, not that there turned out to be no cats, but that cats have turned out not to be animals as we originally supposed. The original concept of cat is: *that kind of thing*, where the kind can be identified by paradigmatic instances. It is not something picked out by any qualitative dictionary definition. (Kripke, 1980, 122)

In contrast, the basic hybrid view holds that if cats were named 'cats' with the help of a trivial description such as 'that kind of thing', the reference fixing must've been

radically indeterminate, as there seems to be no reason the reference would've been grounded in the animal species as opposed to some other general class that cats share in (e.g., felines, mammals, vertebrates) or even a spatio-temporal slice of some cats. On the other hand, supposing that the associated description was 'that animal species', if the creatures called 'cats' turn out to be a species of demons, then the conclusion must be that there are no cats.

I have now summarised the main motivations for the basic hybrid account, which also explain why it's shaped as is. To end this section, I will mention two reasons to be pulled away from the basic hybrid account despite its promise in solving two difficult problems of reference fixing. These reasons are what also motivate the pure causal-historical account of reference fixing.

The first reason is that the original intuition (or inclination, as Kripke put it) to say that even if certain pets of ours turned out to be demons, we would not say that they are not cats, holds some appeal. This inclination is strengthened by the fact that more than a few actual times speakers have been in radical error about the kind they have referred to, yet at the same time it seems that the speakers were not merely speaker-referring[4] to the tokens of the kind but to the kind itself. Richard Miller writes:

> [S]urely medieval European natural scientists could refer to magnets despite being very much mistaken about their underlying nature. Indeed, how could they be mistaken about magnets unless they actually referred to them and had beliefs about them? (1992, 427, fn.)

The second reason to be dissatisfied with the basic hybrid account is that it inherits some of the general problems of descriptivism. (For descriptivism, see Kripke (1980) and a comprehensive discussion by Raatikainen (2020).) I will mention one such inherited problem in particular: incompleteness. Suppose that in successfully naming cats as 'cats' the speakers were required to think of them as an *animal species*, according to the basic hybrid account. (Alternatively, if the description relies on the conjectured common cause of a set of properties of the named samples, then the description of at least some of the properties and/or of the common cause must be true for the reference to succeed.) Now, how did the general term 'animal species' acquire its reference? To avoid circularity, this must have happened independently of naming any animal species. Furthermore, we may assume that the successful reference fixing of 'animal species' itself required some higher-order description in order to exclude alternative, possible reference classes, which then reintroduces the problem at a higher level. Briefly, if any successful case of reference fixing requires a true description, where do the initial descriptions come from? A trivial description will not do here since, as we saw, by definition, a trivial description does not rule out any possible reference class.

There is another tract by which to deepen this argument. It is an observed fact that children learn kind terms before they learn words for higher taxa: 'dog' is learned before 'animal', 'car' before 'vehicle'. Moreover, this is apparently not a simple effect of word-exposure (people around the child talk more about cars than they talk

[4] For the difference between speaker's reference and semantic reference, see Kripke (1977).

about vehicles), 'but children also find it more natural to categorise a novel object as an instance of a basic-level kind than as an instance of a superordinate kind' (Bloom, 2002, 90). Now, if ontogeny is any clue to phylogeny, then it can be conjectured that first speakers of human language could refer to natural kinds in their environment before they could describe these kinds according to *any* higher taxa, never mind the *correct* ones, or even describe their properties. Naturally, this tract is less secure and contains a lot more complications than the actual examples which I presented above for the simple reason that we do not know how language originated, but I think it says a lot to observe that the hybrid theorist must hold that *nowhere* can the reference of a general term succeed unless it is accompanied by a true description of the kind.

This is a difficult puzzle that may or may not be solvable for the hybrid account, among other issues. So, *prima facie*, it gives a reason to take a second look at the pure version of the causal-historical account of reference fixing, which I shall do below.

3 Argument for the pure causal-historical theory of reference fixing for natural kind terms

I'll start with a caveat about the scope of the argument. As already mentioned, the pure causal-historical account of reference fixing claims that some, but not necessarily all, natural kind terms can have their reference fixed by a purely causal grounding. In general, this is not the case with highly theoretical terms like 'electron'. The kinds of cases where pure causal-historical reference fixing is possible include macroscopic observable natural kinds like the elephant, the kangaroo, gold, water, etc. Those are the cases where the following argument applies, though I won't specify this at every turn.

To put the present problem somewhat differently, imagine a speaker perceiving an instance of some macroscopic observable natural kind, say a cat. The speaker points at the cat and says, 'I name that kind of thing "cat"'. The questions are, are there any facts, aside from the descriptions associated in the speaker's mind with the cat, that determine whether the term 'cat' in the speaker's use becomes grounded in the kind cat, as opposed to (a) the kinds of feline, mammal, vertebrate, etc. or (b) one or more of the cat's properties, like being whiskered, being a predator, being cute, etc. Another way to approach the problem is, is there any way to understand Putnam's (1975) same_{L} relation without involving any associated descriptions in the mind of the speaker?[5]

The main pivot of my solution is that 'cat' becomes causally grounded in the property of being a cat (or cathood), as opposed to a higher taxon or another property of cats, because being a cat explains why the cat (a) belongs to the higher taxa that it in fact belongs to (but not the other way around) and (b) why it has the other properties that it actually has (but not the other way around). In a slogan form, cathood is the causal anchor of the grounding act.

Important to this pivot is the concept of 'super-explanatory property' recently coined by Marion Godman, Antonella Mallozzi, and David Papineau (2020). Briefly,

[5] Although I use his terminology, my idea of the same_{L} relation will come to differ from Putnam's.

a super-explanatory property of a natural kind is its property which explains a lot of other properties of the kind as well as why these properties are instantiated together in tokens of the kind. For example, in the case of the natural kind water, the super-explanatory property is its molecular structure; this is what explains (perhaps conditionally on laws of nature) many of water's typical features like its boiling point, freezing point, why it is a powerful solvent, and arguably even why it is necessary for carbon-based life forms. In the case of cathood, the super-explanatory property is more complicated. Unlike with chemical kinds, with biological kinds the super-explanatory property (perhaps better, super-explanatory feature) is not intrinsic but extrinsic, at least according to Godman, Mallozzi and Papineau: it is the phylogenetic history of the cat that explains why its different properties are instantiated together (2020, 325).

Now, let's start the argument with question (b). What is the super-explanatory property of cathood such that it fixes the reference of 'cat' in the speaker's use? On the one hand, the property in question is the cat's genomic material, on the other it is the cat's phylogenetic history, conditional on its growing environment.[6] Together, these things causally explain quite many of the cat's properties, for example why it has whiskers, why it is a predator, and arguably also why it is cute. In contrast, the properties of having whiskers, being a predator, and being cute do not explain why the creature is a cat. A tiger cub would also have these properties without being a cat. In other words, there is a causal structure among the properties of the cat such that having a certain genomic material, a certain phylogenetic history, and the right growing environment explain why the cat has whiskers, is a predator and is cute but not the other way around.

Let's continue to question (a). Why is it that 'cat' comes to be fixed in the property of being a cat as opposed to being a feline, a mammal, a vertebrate, etc. even though the cat is all these as well? The reason is that being a feline, a mammal, or a vertebrate do not causally explain all the common properties of the cat, such as having whiskers, being a predator and being cute, while being a cat does explain them.

How do these observations help to solve the two *qua*-problems? The key is to see the grounding act (speaker pointing at the cat and saying 'That is a cat') as an effect, the cause of which are the perceived properties of the cat, such as having whiskers, being a predator and being cute. The properties in turn are causally structured so that their mutual coexistence in the cat is explained by the property of being a cat, which in important part consists of having a certain genomic material and a certain phylogenetic and ontogenetic history. In this relation, being a feline, a mammal, a vertebrate or an animal do not play a super-explanatory causal role. Therefore, they are excluded from being the referent of 'cat' by the pure causal-historical account. This can be seen in that if the creature turned out not to be an animal, but a species of demon, the reference would still be the cathood of the cat, only cathood's super-explanatory property would differ from what we assumed.

[6] This is somewhat different from how Godman, Mallozzi, and Papineau use the term 'super explanatory property' with biological kinds: for them, the phylogenetic history plays the key super-explanatory part. I have opted for a rougher use because my main argument can be made independently of the fine details and commitments around the notion of super-explanatoriness in the case of biological kinds, which are contested in the literature.

Another way to phrase this argument is that in the case of the cat, Putnam's $same_L$ relation turns out to be a causal, non-descriptive structure. Imagine the speaker encounters another animal that looks a lot like the cat she met earlier but which lacks whiskers. What determines whether this animal is of the same relevant kind as the first one, i.e., a 'cat'? The answer is that the new animal is a 'cat' if it shares (roughly) the same genomic material and phylogenetic history as the first 'cat', in other words, if the super-explanatory property is the same. (This is, of course, only to provide a rudimentary picture how actual species relations are identified, but that is not a problem unique for the causal-historical account.)

It is noteworthy that this is not how Putnam, at least in *The Meaning of Meaning*, understood the $same_L$ relation. For him, our interests (or the context) in counting things as being of the $same_L$ also mattered:

> For example, I might say "did you see the lemon," meaning the plastic lemon. A less deviant case is this: we discover "tigers" on Mars. That is, they look just like tigers, but they have a silicon-based chemistry instead of a carbon-based chemistry. (A remarkable example of parallel evolution!) Are Martian "tigers" tigers? It depends on the context. (1975, 157–158)

Here, my goal is not to refute Putnam's understanding of the $same_L$ relation, but to argue that there are other ways to understand it that do not include our interests or associated descriptions. That being said, in the case of 'Martian tigers' the judgement of pure causal-historical account would differ since Martian tigers would not share in the same super-explanatory property as actual tigers. Hence, they would not count as the same species, even if for our zoo-keeping interests it would be just as well to have Martian tigers on display.

In response to the argument above, an anonymous reviewer made two objections that are worth bringing up here. The first objection is that being super-explanatory does not explain why the term 'cat' is fixed on cathood, as opposed to the properties of being a feline, a mammal, a vertebrate, an animal, etc., because these latter properties also are super-explanatory, only of different properties than cathood. So, why is reference fixed on one super-explanatory property as opposed to another? The second objection is that my proposal entails that it is impossible to ground the reference of terms to taxa higher than species purely causally, which would be an unwelcome consequence.

To answer these objections, it is good to have a second look at what it means for a property to count as super-explanatory. Godman, Mallozzi, and Papineau state that for them, a super-explanatory property explains *all* the common properties of the kind, although they leave it open whether this is a necessary and/or a sufficient criterion for being super-explanatory (2020, 320). However, I think that a property can be counted as super-explanatory, at least for the purposes of my argument, if it explains the vast majority of the common properties of the kind, especially readily perceived properties.

With this adjustment of super-explanatoriness in mind, let's take a second look at the example of the cat. The answer to the first objection is that the reference of 'cat' is fixed on the property of cathood because this property is super-explanatory *relative to*

the set of properties (readily) perceived by the speaker during the naming act. In contrast, while the properties of being a feline and a mammal do explain some properties of the cat, they do not explain the vast majority of its properties, while being a cat does explain the vast majority. Hence, being a cat is the only super-explanatory property around, so there is no competition between different super-explanatory properties relative to the set of readily perceived properties.[7]

This response also contains the key to resolving the second objection. The important point is that being super-explanatory must be understood as relative to the set of properties perceived by the speaker during the naming act. In order to ground the term 'animal' to the property of being an animal purely causally, the speaker must perceive a set of properties such that being an animal is super-explanatory relative to them. A natural story of how this might happen is that the speaker perceives a set of different animal species, each of which she calls 'animals'. (Observe that the hybrid theorist takes the opposite direction: in order to fix the reference of 'cat', the speaker must already have fixed the reference of 'animal species'– or some other term – because she uses this term referentially in her associated description of the cat.)

The first objection can be pressed on further by rephrasing it. Consider the set A, which includes all the perceivable properties of a token cat. A subset of A, C, includes all the perceivable properties which are super-explainable by the cathood of the cat. A subset of C, F, includes all the perceivable properties that are super-explainable by the felinehood of the cat, and the subset of F, M, includes all the perceivable properties which are super-explainable by the mammalianhood of the cat. So, there are three different properties that are all super-explanatory relative to some properties perceived by the speaker. Why does 'cat' fix its reference on one of them as opposed to others?

The reason is that only the subset C is super-explanatory relative to the total set of perceivable properties A of the token cat. Notice that C need not explain all of A under my adjusted understanding of super-explanatoriness; it suffices that C explains the vast majority of them, unlike F and M. The reason why F and M cannot explain as many properties as C is that they are subsets of C, which simply means that there are some properties of the token cat which cannot be explained by its felinehood or mammalhood, but which can be explained by its cathood. And this is what suffices, I claim, to fix the reference of 'cat' in the super-explanatory feature of cathood. This also should bring more clarity to when a term used by the speaker fixes its reference on a taxon higher than a species: when that taxon plays a super-explanatory role to the total set of perceived properties.

Next, I will briefly compare my argument to Miller's (1992) proposal. My argument is in line with his in the spirit that '[t]he qua is built into the causal power' (Miller, 1992, 429). The cause of the grounding act is the property of the object named which stands in super-explanatory relationship to the object's other (per-

[7] Is it possible for a set of perceived properties of a kind to have more than one super-explanatory property? How is the reference fixed then? While this is an important question concerning further study and use of the idea of super-explanatory properties, I think that for quite many kinds, e.g. the cat, there is only one super-explanatory property around, which suffices to resolve the objection.

ceived) properties.[8] In the case of the cat, this means, roughly, its genomic material, ontogenetic and phylogenetic history. These are what determine when two creatures are both cats because they causally explain the typical properties of cats in virtue of which the speaker perceived one even when two cats don't share all the properties which originally caused the grounding act. The reason why the higher taxa of the cat are excluded from being the referent of 'cat' is that these properties are not super-explanatory in relation to the set of properties perceived by the speaker when naming a token cat as 'a cat'. The speaker would have perceived the creature's whiskers, predatory behaviour, and cuteness even if it had been a demon, in which case cats would have turned out to be demons. On the other hand, the properties of having whiskers, being a predator, and being cute, while causally active in the grounding act, are not the referent because they are not super-explanatory, i.e., they do not explain the other typical properties of cats as well as their co-instantiation. There can be cats which lack whiskers, have lost their predatory behaviour and which aren't even cute. In other words, non-super-explanatory properties do not have the necessary generality to include all the creatures we intuitively think belong to the extension of 'cat'.[9]

My proposal differs from Miller's in that I, like Deutsch (2021), don't see it necessary to include the ability to identify cats as being part of the effect in the causal grounding relation. While cats are relatively easy to recognise once one has seen one, there are exceptions: if the speaker has only seen very hairy cats before, seeing a hairless one might discourage them from calling it a cat. Yet it would still be a cat. Although I'm not pressing the argument here, the main reason that I think the ability to recognise sameness is not so important for fixing reference is that structural similarity can exist in nature even when our ability to recognise it is poor or even non-existent. As I argued above, the relevant sameness comes in the form of a causal structure, which is independent from our ability to identify it. (For more good arguments against the recognitional capacity version of pure causal-historical account, see Gómez-Torrente (2019, 157–158).) That being said, without any reliable capacities to recognise the referents of our terms, empirical language would be pretty useless. But that point differs from the claim that recognitional capacities are needed to solve the *qua*-problems.

Now, my proposal is not a full answer to the *qua*-problem thematic. What fact determines that the speaker isn't naming a spatio-temporal slice of the super-explanatory property of cathood, but cathood as such? Since both the slice and the whole would presumably carry identical effects, the solution remains at large. That being said, so long as the *qua*-problems need not be clumped into one intangible package, there is no reason to believe that another pure causal solution could not be found for

[8] Of course, as an intentional act, the grounding act affords many causal explanations, one of which is that the speaker wanted to name something. But the particular cause that matters here is the cause *qua* which the reference of 'cat' is fixed.

[9] This explanation only holds, of course, if 'cat' actually refers to cats as opposed to, say, a set of descriptive properties that may or may not be true of the creatures called 'cats'. This alternative would effectively amount to a version of descriptivism. While I think this idea is nowadays highly implausible, it's not the aim of this paper to refute it; I refer the reader to Raatikainen (2020) for a recent overview of the challenges of descriptivism.

this problem. Perhaps Deutsch (2021) is right, and the problem will be solved along with the more general problem of causal underdetermination.

To finish this section, I will clarify my use of super-explanatory property in the case of biological kinds. Notice first that it is controversial whether biological kinds should be counted as 'eternal' or 'historical' kinds, or perhaps a mixture of the two. (The vocabulary of historical and eternal kinds is due to Millikan (1999).) Devitt (2008;, 2010) for one has argued that there can be biological eternal kinds with intrinsic essences. If he is right, then biological kinds will have super-explanatory properties in the way chemical kinds do. The main thing is that some properties of the cat, such as its genomic material combined with its phylogenetic and ontogenetic history, together explain most of the cat's other properties, such as having whiskers, being a predator and being cute, but not the other way around. This is what for the most part explains why cats form a kind, i.e., their intrinsic similarities. Therefore, it also explains in what the term 'cat' is fixed in the grounding act. The notion of super-explanatoriness is relatively new, so it remains to be seen how many different kinds of kindhood reference it can be used to explain.

4 On reference failure

Most everyone agrees that sometimes our terms fail to refer despite our best efforts to the contrary. 'Phlogiston' is a parade example. I agree with, e.g., Cummiskey (1992) and Deutsch (2021, 1819, fn.15) that the best that the pure causal-historical account can hope for here, as with highly theoretical terms in general, is to rely on associated descriptions to explain the failure. In this section, I propose a way for the pure causal-historical theory to explain reference failures in the case of non-theoretical terms that is independent of the truth and falsehood of the associated descriptions.

We can start by presenting a thought experiment of reference fixing owing to Amie Thomasson:

> *Key Sparrows.* If ornithologists coin the term 'Key sparrow' to name a new race of sparrows apparently discovered to be living in the Florida Keys, only later to find that all of the supposed exemplars observed were sophisticated animatronics planted by a glory-hungry birdwatcher, it seems we would say that there are no Key sparrows (since the things observed were not birds at all), not that it turns out that Key sparrows are little robots. (Thomasson, 2007, 49)

Deutsch observes that this is not a counterexample to the pure causal-historical theory of reference fixing because that theory allows the possibility that some terms get their reference fixed with the help of a description, which, if false, will mean that the fixing fails (2021, 1818). However, while this response is logically possible, it raises the question of when exactly should an attempt to fix the reference of a non-theoretical term like 'Key Sparrows' be counted as descriptive, when purely causal? In order to test the theory, either in thought experiments or in actual practice, it is important to have a criterion of reference success and failure that is independent of what intuitively the right answer is. Otherwise, the pure causal-historical account risks clas-

sifying examples on an ad hoc basis. This is what Deutsch arguably does when he claims that a very similar thought experiment, called 'Adam and the elephant' (owing to Mallon, 2016), would have resulted in a successful reference fixing even if the creature named turned out to be a robot, despite the fact that the speaker intended to name an animal species as 'elephant' (Deutsch, 2021, 1814).

The general problem is this. *Prima facie*, the examples of Key Sparrows and Adam and the elephant seem very much alike. So, we should expect them to behave alike in terms of success and failure conditions. Yet some see an intuitive difference here. Personally, I think 'Key Sparrows' is a case of successfully fixing the reference to the animatronics, which is also suggested by Dodd's (2012) discussion, while Deutsch and Thomasson disagree. Moreover, even if 'Key Sparrows' can be counted as a successful reference fixing, there are some non-theoretical terms which certainly cannot be, like 'witch'. Is this always due to an associated description having turned out to be false? I think not.

In brief, my proposal is the following. Reference fixing can fail both for theoretical and non-theoretical terms. In the case of theoretical terms, the failure is due to an associated description turning out to be false. This is why 'phlogiston' does not refer. But in the case of non-theoretical terms, the nature of referential success and failure is something else than truth and falsehood of the associated descriptions. In the non-theoretical case, success means that there is some super-explanatory property that in fact explains the same_{L} relation among tokens of the named kind. In contrast, failure means that there is in fact no super-explanatory property that explains sameness in the causal structure of the candidate tokens. I'll illustrate both sides, first with the help of Key Sparrows, second with the help of witches.

It may seem intuitive that the reference of the general term 'Key Sparrows' fails because reality does not match the descriptive intentions of the speakers. However, this is compatible with saying that although the speakers failed in their intention to name a new species of sparrows, they succeeded in naming a new kind of animatronics instead. In general, there is no conflict in succeeding in doing one unintended thing while failing to do another, intended thing, as Deutsch also observes (2021, 1814). So, the naming of animatronics as 'Key Sparrows' is accidental. After all, it is plausible that, once informed of the reality, the speakers decide to retain the term 'Key sparrow' as a name for the animatronics: this is suggested by the critical discussion of Dodd (2012). The only way in which this example is objectionable to the pure causal theory is if it is insisted that the reference grounding of 'Key Sparrow' failed already at the time of the grounding act, so that the descriptive intention rules out the possibility of correctly using 'Key Sparrow' as a term for the animatronics in the future, absent a new grounding act with a different descriptive intention. This is to say that the intention to describe the animatronics as sparrows failed, and that this was *all* there was to the grounding act, that it wasn't *also* an act of grounding the term in the animatronics. But why should we think it was not? To reject this reading, I think it is the burden of the hybrid theorist to say why, in this case, it is not possible to fail in one intended action and simultaneously succeed in another unintended action, which in general seems perfectly possible.

The case of Key Sparrows is starkly different from the case of witches because for witches, there simply isn't anything real in the causal structure of candidate witches

themselves that would explain their being of the same, distinct kind. The animatronics belong to a distinct (non-natural) kind in virtue of being artefacts with the same maker, design, materials, and function, whereas 'witches' were simply unlucky individuals and victims of prejudice. So, the idea is that there is a super-explanatory property, analogical to chemical structure and phylogenetic history, that explains the common properties of the animatronics, considered as an artefactual kind.

Naturally, most of the associated descriptions for witches were also false. But this alone does not for certain explain why the reference failed. Imagine, for instance, that some or most individuals called 'witches' in fact had strange, apparently magical powers, but these were not due to a pact with the Devil but rather to Martian technology which they had found and put to their benefit, forming a cult to protect their secret. In such a case, would we still say that witches were not real? Or would we instead say that they were real, only that their powers were not due to the Devil but rather to Martian technology? If the answer is the latter, then the reference of 'witch' isn't dependent on the truth of the associated descriptions. That being said, while thought experiments like this can be useful in testing our theories of reference, we should be cautious about putting too much credence in their power to settle disputes. The main reason 'witches' doesn't refer is that there just isn't any distinct (natural or non-natural) kind which candidate witches shared in, including no super-explanatory property.[10]

Although candidate witches do not actually form a kind in the same sense that cats and animatronics of a given design do, it is still possible that there were some real similarities between victims of witch trials, in each historical context. Most of them were women, for instance. So, the question arises, how much structural similarity is enough for reference to succeed, and how little to fail? I don't know if there is any general answer to this difficult question. Nonetheless, the point of this brief discussion was to sketch a strategy for the pure causal-historical account to deal with reference failure in the case of non-theoretical terms. The key is to look, not for the truth of the associated descriptions, but actual causal structures and similarities in the tokens themselves, some of which may turn out to be surprising, as in the case of Key Sparrows. But even when such surprising structures can be found, we may still end up saying (within a margin of certainty) that 'Key Sparrows' fails to refer simply because we decide how we want to use our words, and signalling that a given term is regarded as empty is a good way to abandon using it.

[10] There is one distinct kind to which arguably all candidate witches belong(ed) to, namely the social kind of being treated as a witch. The idea might seem obviously wrong-headed, yet I do think there is something interesting to be said on its behalf, assuming it is starkly distinguished from other cases of finding 'surprising' referents like with Key Sparrows. Briefly, the case of witches could be viewed as analogous to debates over the reference of race and (possibly) gender terms. Although race is not a category recognized by the current science of human biology, thus not a distinct natural kind, racial terms are and have been widely used referentially. One answer to the apparent puzzle is that racial terms refer to racialized human groups (Haslanger, 2012, Ch.10). Similarly, one could think of a society which believed in hereditary witchcraft, and whose classifying practices created the referent for 'witch'.

5 Conclusions

The main claim of this paper is that the reference of some general terms can be grounded in purely causal fashion, meaning without mediation by any true descriptive, non-trivial intentions or beliefs. I argued for the theoretical coherence of the pure causal view by deepening the description of the referential relation, originally made, among others, by Miller (1992) and Deutsch (2021), as a causal relation with the help of Godman, Mallozzi and Papineau (2020). The key idea is to look for a property which explains the vast majority of the common (perceived) properties of the kind. Different kinds may turn out to have different kinds of super-explanatory properties. For chemical kinds, this means their microstructure; for biological kinds, their ontogenetic and phylogenetic history; for artefactual kinds, their maker, design, function, and materials. Further study is needed to show the true potential of this notion.

Acknowledgements I warmly thank all the anonymous reviewers for their encouraging and insightful comments, which led to significant improvements of the argument and readability of the paper.

Funding Open access funding provided by Tampere University (including Tampere University Hospital).

Declarations

Ethics approval and consent to participate This article is not subject to conflicts of interest or to any other ethical issues known to the author.

References

Bloom, P. (2002). *How children learn the meanings of words*. MIT Press.

Cummiskey, D. (1992). Reference failure and scientific realism: A response to Meta-induction. *The British Journal for the Philosophy of Science*, *43*(1), 21–40.

Deutsch, M. (2021). Is there a 'Qua Problem' for a purely causal account of reference grounding? *Erkenntnis*. https://doi.org/10.1007/s10670-021-00428-3.

Devitt, M. (1996). *Coming to our senses: A naturalistic program for semantic localism*. Cambridge University Press.

Devitt, M. (2008). Resurrecting biological essentialism. *Philosophy of Science*, *75*, 344–382. https://doi.org/10.1086/593566.

Devitt, M. (2010). Species have (partly) intrinsic essences. *Philosophy of Science*, *77*, 648–661.

Devitt, M. (2015). 'Should Proper Names Still Seem So Problematic?' In Andrea Bianchi (Ed.), *On Reference* (Oxford, 2015; online edn, Oxford Academic, 23 Apr. 2015). https://doi.org/10.1086/656820.

Devitt, M., & Sterelny, K. (1999). *Language and Reality: An Introduction to the Philosophy of Language* Second Edition, Blackwell.
Dodd, J. (2012). Defending the Discovery Model in the ontology of art: A reply to Amie Thomasson on the *Qua* Problem. *'British Journal of Aesthetics*, *52*(1), 75–95.
Godman, M., Mallozi, A., & Papineau, D. (2020). Essential properties are Super-explanatory: Taming metaphysical modality. *Journal of the American Philosophical Association*, *6*(3), 316–334. https://doi.org/10.1017/apa.2019.48.
Gómez-Torrente, M. (2019). *Roads to reference: An essay on reference fixing in natural language*. Oxford University Press.
Haslanger, S. (2012). *Resisting reality: Social Constructionism and Social Critique*. Oxford University Press.
Kripke, S. (1977). 'Speaker's Reference and Semantic Reference.' *Midwest Studies in Philosophy*, II.
Kripke, S. (1980). *Naming and necessity*. Blackwell.
Mallon, R. (2016). *The construction of human kinds*. Oxford University Press.
Mallon, R. (2017). Social Construction and Achieving Reference. *Noûs*, *51*, 113–131. https://doi.org/10.1111/nous.12107.
Miller, R. (1992). A purely causal solution to one of the qua problems. *Australasian Journal of Philosophy*, *70*(4), 425–434.
Millikan, R. (1999). Historical kinds and the 'Special sciences'. *Philosophical Studies*, *95*, 45–65.
Putnam, H. (1975). 'The Meaning of Meaning'. *Language, mind, and knowledge. Minnesota studies in the philosophy of science*, Volume 7 (1975), 131–193.
Raatikainen, P. (2020). Theories of reference: What was the question? In A. Bianchi (Ed.), *Language and reality from a naturalistic perspective* (Vol. 142). Springer. Philosophical Studies Series.
Stanford, P. K., & Kitcher, P. (2000). Refining the causal theory of reference for natural kind terms. *An International Journal for Philosophy in the Analytic Tradition*, *97*(1), 99–129.
Thomasson, A. (2007). *Ordinary objects*. Oxford University Press.