Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Folk Intuitions about Reference Change and the Causal Theory of Reference

STEFFEN KOCH
Institute for Philosophy, Bielefeld University
&
ALEX WIEGMANN
Institute for Philosophy II, Ruhr University Bochum

In this paper, we present and discuss the findings of two experiments about reference change. Cases of reference change have sometimes been invoked to challenge traditional versions of semantic externalism, but the relevant cases have never been tested empirically. The experiments we have conducted use variants of the famous Twin Earth scenario to test folk intuitions about whether natural kind terms such as 'water' or 'salt' switch reference after being constantly (mis)applied to different kinds. Our results indicate that this is indeed so. We argue that this finding is evidence against Saul Kripke's *causal-historical view* of reference, and at least provisional evidence in favor of the *causal source view* of reference as suggested by Gareth Evans and Michael Devitt.

Keywords: semantic intuitions, natural kind terms, reference change, Kripke, causal-historical view, causal source view

1. Introduction

Descriptivism holds that the links between proper names and their referents are established by the definite descriptions which speakers associate with them (e.g., Frege 1948; Russell 1905; Searle 1958). Kripke (1980) challenged descriptivism and proposed what became known as the *causal-historical view* of reference. On this view, the reference of a proper name is fixed by ostension or a reference-fixing description. Once reference is fixed in this way, it is preserved through communicative chains. Importantly, this does not require the speaker to know how the term was used before; the mere *intention* to use it as it has previously been used suffices for successful reference.

Contact: Steffen Koch <steffen.koch@uni-bielefeld.de>
Alex Wiegmann <alexander.wiegmann@ruhr-uni-bochum.de>

https://doi.org/10.3998/ergo.2226

Kripke argues that a view along these lines not only applies to proper names but also to natural kind terms. This requires supplementing the causal theory of reference with a view of natural kinds. According to Kripke's natural kind essentialism, natural kinds like water and gold are constituted by their essential properties. Consequently, he argues, natural kind terms denote all and only instances of the kind to which the term was fixed via introduction. For instance, when a speaker introduces the term 'water' to refer to watery stuff like that before her, then all and only those entities which share the essential properties of this liquid stuff will fall within the extension of 'water'. Again, speakers, including those who introduce a term, need not know what these essential properties are—the mere intention to use the term in the same way in which it was introduced is sufficient.[1]

Largely due to Kripke's and Putnam's enormous influence, descriptivism became rather unfashionable. However, in the past two decades, Kripke-Putnam-style externalism has come under attack from experimental philosophy. A number of independently conducted studies seem to show that a significant minority of lay people respond to the relevant thought experiments (e.g., the Gödel case or the Twin Earth scenario) in ways that are predicted by descriptivist rather than externalist views (e.g., Machery et al. 2004). Furthermore, it seems that people's intuitions about such cases vary on the basis of presumably irrelevant influences such as *culture* (Machery et al. 2004; 2010; 2015; Beebee & Undercoffer 2015; 2016; Sytsma Livengood, Sato, & Oguchi 2015) or *gender* (Buckwalter & Stich 2015).[2] In light of these findings, it is now less clear how robust the externalist case judgments really are.[3]

However, descriptivism is not the only rival to the Kripkean view. There is also an important but often neglected in-house dispute between proponents of different broadly causal theories of reference. On Kripke's own view, the causal-historical view, descriptive content is not necessary for successful reference. Although they share (and predict) the Kripkean case judgments against descriptivism, Evans (1973) and Devitt (1981) propose what we will call the *causal source view* of reference. Just like Kripke's view, this view takes reference to be established by causal relations rather than semantic fit between an associated description and the referent; but unlike Kripke's, this view does not take the relevant causal relation to hold between the original act of baptism and the term's

1. For discussion and criticism of natural kind essentialism, see, e.g., Koslicki (2008), and Häggqvist and Wikforss (2015).

2. But see Adleberg, Thompson, and Nahmias (2015) and Seyedsayamdost (2015) for failed replications of the gender effects reported by Buckwalter and Stich.

3. See Braisby, Franks, and Hampton (1996), and Jylkkä, Railo, and Haukioja (2009) for more empirical testing of the Kripkean view of natural kind terms; see Häggqvist and Wikforss (2015) for discussion. See ch. 2 of Machery (2017) for a useful summary of the existing data.

current use, but rather between the item's properties and the mental content that the subject associates with the term in question. The resulting view has it that descriptive content does, after all, play a role in reference determination. But in contrast to descriptivist views, the referent is not determined by the *content* of the description, but rather by what happens to be its *causal source*. Devitt and particularly Evans argue that the causal source view fares a lot better than the causal-historical view when it comes to cases of reference change. Both authors put forward a number of cases in which a term intuitively switches reference in ways that their views would predict but which are left unexplained by the causal-historical view.[4]

The cases used by Evans and Devitt to argue for a modification of Kripke's original view have never been tested empirically. This paper fills this gap by presenting the results of two experiments we conducted on such cases. In these experiments, lay people were asked to assess whether a natural kind term switches reference despite the language users having the intention to preserve the original reference. The results of our experiments provide evidence against Kripke's causal-historical view, and support causal source views along the lines of Evans's and Devitt's (with one caveat: see Section 4). They also show that the typical externalist case judgments—those shared by Kripke, Evans and Devitt—are more widespread (at least among native English speakers) than previous studies suggest. Taken together, this gives experimental support to the causal source view of reference (at least with respect to native English speakers).[5]

The structure of the paper is as follows. In Section 2, we sketch both the causal-historical view and the causal source view of natural kind terms and explain why, despite apparent similarities, they are committed to diverging case judgments in cases of alleged reference change. In Section 3, we state the results of our (preregistered) online experiments, showing that people's truth value judgments support the causal source view rather than the causal-historical view. In Section 4, we discuss the methodology and results of our experiments in more detail, also indicating an important caveat concerning the interpretation of our data.

4. See also Fine (1975) for this line of critique against Kripke's and Putnam's views about reference.

5. Tobia, Newman, and Knobe (2020) recently investigated how people categorize natural kinds. An interesting finding of their study is that people's judgments about natural kinds exhibit a dual character pattern: they are not driven by just one, but instead by two separate sets of criteria. These different sets of criteria generate opposing verdicts in Twin Earth cases, hence the complexity and ambivalence of how people react to them. It would be interesting to investigate the philosophical significance of these findings with respect to causal theories of reference. However, since the main task of this paper—to investigate intuitions about reference change and to use them to adjudicate between two versions of the causal theory of reference—is largely orthogonal to the dual character hypothesis, this will have to wait for another occasion.

2. The Causal-Historical View, the Causal Source View, and Reference Change

Kripke's causal-historical view of reference, formulated for natural kind terms, has the following two elements:

> *Reference-Fixing*. The reference of a natural kind term *t* is established via ostension or a reference-fixing description of a set of paradigm instances of a kind *k*. A mechanism of implicit or explicit generalization makes sure that all and only tokens of *k* are in the extension of *t*. Whether or not a token belongs to *k* is a matter of whether the token has the relevant shared essential properties.
>
> *Reference-Preservation*. Once reference is established, it can easily be transmitted via chains of communication. For a speaker to gain the ability to refer to a kind *k* with the term *t*, it is sufficient that she hears someone use *t* correctly, and that she intends to use *t* just as this person did.

Kripke presents the above view as a "picture" rather than a theory (Kripke 1980: 93). As a consequence, not all of the notions he uses to describe this picture are precisely defined. This is mostly unproblematic for our purposes. However, one thing that needs clarification is the notion of an intention. Kripke writes that a speaker "must . . . intend when he learns [a term] to use it with the same reference as the man from whom he heard it" (1980: 96). There is a *de re* and a *de dicto* reading of this passage. On the *de re* reading, a person must have *the same intention* as the person from whom she heard the name, that is, the intention to refer to a particular object or person. On the *de dicto* reading, a person must have the intention to use the term *in the same way* as this other person did (however that was). However, both of these readings pose far too demanding conditions on successful reference to be consistent with Kripke's overall view. We therefore opt for a minimal reading, according to which it is generally sufficient for successful reference preservation that a person *does not* have the intention of using the term differently than those from whom she heard it.

Among those who are generally sympathetic to causal views about reference, some have raised concerns about Kripke's particular implementation of it. In this vein, Gareth Evans (1973) argues that Kripke's view is unable to accommodate the phenomenon of a name changing reference. But since cases of reference change seem not only possible but actual, Evans thinks of such cases as counterexamples to the Kripkean view. These cases share a common structure: First, a name is properly fixed to refer to an individual. Then someone mistakenly applies the term to a new individual. Others go on borrowing the term from the transgressor. Over time, the new usage becomes so dominant that it seems the name has permanently switched reference. Evans illustrates this phenomenon

with two example cases: the case of 'Madagascar', which once referred to parts of the African mainland but now (apparently) refers to the East-African island (even though, supposedly, nobody had the intention of breaking the original reference); and the case of 'Jack' and 'John', the names of two accidentally switched babies which now (apparently) have correspondingly switched reference.

The close analogy which many externalists see between proper names and natural kind terms suggests that something similar could happen for natural kind terms as well (Fine 1975; Koch 2021a). Suppose the term 'water' was introduced to denote a set of paradigm instances of water, and everybody in the language community has the intention of using the term accordingly, although nobody knows about the chemical constituents of water. Now a group of people go to Twin Earth, an uninhabited planet with a liquid indistinguishable from water except that it consists of XYZ rather than H_2O. If the group of Earthlings nevertheless continue calling this substance 'water', then after some time what 'water' refers to will switch from (merely) H_2O to (also) XYZ.

Notice that the case description above implies that there is something wrong with the causal-historical view. On this view, a proper introduction plus the intention to preserve the original reference by later speakers is sufficient to preserve reference. But both of these conditions are satisfied in the above cases: 'Madagascar' referred to parts of the mainland, and nobody had the intention of changing its reference, but nevertheless it did change. The same holds, *mutatis mutandis*, for the other cases.[6]

Evans uses these cases to motivate a different kind of causal theory of reference. As he puts it, the problem with Kripke's causal-historical view is that Kripke "has mislocated the causal relation; the important causal relation lies between [an] item's states and doings and the speaker's body of information—not between the item's being dubbed with a name and the speaker's contemporary use of it" (Evans 1973: 197). This yields the following view about reference determination: The reference of a proper name/natural kind term is "the source of causal origin of the body of information that S has associated with it" (1973: 197.). A similar view is also proposed by Devitt, according to whom the reference of a proper name/natural kind term as used by a speaker S is the object that grounds those thoughts of S which dispose S to use the term (Devitt 1981).[7]

6. To be fair, Kripke acknowledges the possibility of reference change and briefly considers the explanation that "a present intention to refer to a given entity . . . overrides the original intention to preserve reference in the historical chain of transmission" (Kripke 1980: 163). But since he does not specify the conditions for such "overriding" to take place, we do not know how to incorporate this idea into his general view of reference determination. Indeed, Kripke admits that reference change "requires more apparatus than [he has] developed here" and therefore suggests leaving the problem for further work (1980: 163).

7. Evans states his view only with respect to proper names, not natural kind terms. But the view can be generalized from names to natural kind terms in the usual externalist manner (Devitt 1981: esp. ch. 7; Devitt & Sterelny 1999).

As it stands, the causal source view leaves many details undeveloped. In this respect, it resembles the rather pictorial nature of Kripke's articulation of the causal-historical view. For instance, Evans's view is complicated by the fact that people will typically associate all sorts of different information with a name—information that may easily be causally derived from different individuals or kinds. Devitt's view faces a similar challenge. Speakers' dispositions to use terms will often be grounded in different objects or kinds of objects. To account for this, Evans introduces the notion of dominance, and states that reference is determined by what happens to be the dominant causal source of the information that one associates with a name (Evans 1973: 199f.). Although this idea seems plausible, it needs to be spelled out in more detail. Evans notes that dominance is not a purely quantitative issue and mentions other factors that will be involved, but his remarks hardly amount to a worked-out theory (1973: 201). Moreover, if one is to apply the causal source view to natural kind terms, one needs a plausible answer to the qua problem—the problem that any given object instantiates many different natural kinds and that it is thus unclear which of them is selected (cf. Devitt & Sterelny 1999; LaPorte 2004; Thomasson 2007). Dickie (2015: ch. 5.2) also argues that even a refined version of Evans's view faces difficulties in accounting for the intuitive judgments about some particular cases. An overall defense of the causal source view requires addressing these and potentially further intricacies. The aim of this paper, however, is more modest, namely to compare whether its predictions about cases of (alleged) reference change are more intuitive than those made by the causal-historical view. For this reason, we can set these further intricacies aside and operate with the simple version of the view stated above.

Unlike the causal-historical view, the causal source view allows for unintentional reference change in cases like the above. Using the example of 'water', this might work as follows. When the Earthlings go to Twin Earth, all their 'water'-associated beliefs are causally grounded in water, that is, H_2O. So, when they arrive on Twin Earth, their 'water'-involving utterances will be about water, not twin water. However, as time passes, more and more of the information they associate with 'water' will become grounded in twin water, as all the experiences they have on Twin Earth involve twin water rather than water. For this reason, the reference of 'water', in their mouths, will shift from excluding XYZ to including it, and eventually perhaps even to excluding H_2O altogether.[8]

8. Although the above cases are all instances of unintentional reference change, Koch argues that the causal source view works as a model for intentional reference change as well and thus provides a useful metasemantics for conceptual engineering. See Koch (2021b) for a detailed exposition of the view; see Koch (2021a; 2021c) for arguments to the effect that the causal source view gives us some degree of meaning control.

Evans (1982) discusses a hypothetical case of reference change that adds important detail to this rather schematic outline (see also Dickie 2015: ch. 5.2 for discussion). In particular, Evans mentions three different phases of reference change. In phase 1, an individual or entity *x* is switched with *y* without anybody noticing. Evans remarks that "[a]t this early point . . . the name 'NN', as used by anyone who participates in this practice, still refers to *x*" (Evans 1982: 388). In phase 2, we suppose that the substitution of *x* and *y* goes unnoticed and that *y* is continuously recognized by the community as NN. Information that is causally derived from *y* is now disseminated among the members of the community. At the same time, however, some information that is derived from *x* remains. "At this point", Evans argues, "we can say that the name 'NN' . . . no longer has a referent. The persistent identification *of y* as NN has undermined the connection which tied the name uniquely to *x*" (1982: 389).

Evans notes that, in theory, we can imagine a phase 3 in which the name 'NN' has unambiguously switched reference to *y*:

> there is no theoretical obstacle to the loss of all information derived from *x*; and when this happens, the name may finally be regarded as a name *of y*. There is a group of speakers who call *y* 'NN'—who know *y* as NN—and no group of speakers who know anything else as NN is relevant any longer. (1982: 390)

However, it is important to stress that reaching phase 3 will typically be hindered by practical hurdles, because it is so very hard to erase all, or even close to all, information that is derived from *x* and which people associate with 'NN' from a community, especially when people are not aware of the fact that *x* and *y* are not identical. Evans argues that even if we suppose a gradual replacement of the members of the community who knew *x* by members who only know *y*, we should still expect ambiguous reference, because

> there may remain . . . a good deal of information derived from *x*, and these are traces . . . of using the name to refer to *x*. So long as any serious quantity of these traces remains, the practice can still reasonably be said to embody a confusion. (1982: 390)

Reference change, according to this view, is messy, and dependent on numerous contingent factors, such as people's memory and the demographics of the community. So although the causal source view predicts a significant change of reference in the direction of the new entity, it does not predict that this entity is the unambiguous new referent of the name after some fixed period of time.

Here we have two different versions of the causal theory of reference: Kripke's causal-historical view and Evans's and Devitt's causal source view. Among other things, the two views differ in how they treat cases such as the above. Whereas the causal-historical view takes reference to remain rather stable in such cases, the causal source view predicts that such cases initiate a process of reference change that may, at least in principle, lead to unambiguous reference change (although it remains unclear when this point is reached). Importantly, all three authors—Kripke, Evans, and Devitt—base their views on case judgments (e.g., Kripke's Gödel case, Evans's Madagascar case, or Devitt's grugru case). According to the majority view, these case judgments are supposed to be shared by lay people. If it turns out that one view's predictions are better supported by the judgments of lay people than those made by an opposing view, this constitutes (defeasible) evidence in favor of the former. One way to adjudicate between the two versions of the causal theory of reference is therefore to test empirically which of their predictions about cases are shared by language users.[9]

One caveat before we start: A further upshot of the pictorial nature of Kripke's view is that it is difficult to assess. Necessary and sufficient conditions can be falsified by counterexamples. But how does one argue with pictures? Our strategy to deal with this problem will be as follows. We take Kripke's articulation of his view in *Naming and Necessity* seriously, and more or less ignore his remark about it being a picture rather than a theory. Should Kripke's picture indeed be inconsistent with the data we will state and discuss later on, then the central claims of this paper go through even though Kripke might not be committed to the exact wording of the view we suggest above. We take this to be the likely outcome. But should his view be flexible enough to accommodate the findings offered here, then this paper should be seen as an argument for exploiting this flexibility and interpreting him in a way that is consistent with the evidence we shall provide.

3. Experiments

The goal of our experiments was to find out which of the two views—the causal-historical view or the causal source view—is better supported by folk intuitions about cases like those above. In particular, we ran two structurally analogous

9. That being said, there remains some controversy over how philosophers actually support their preferred case judgments. Whereas a majority think that case judgments are usually based on intuitions, some claim that they are instead based on arguments (cf. Cappelen 2012; Deutsch 2015; Horvath & Koch 2021). We will not delve into this large metaphilosophical debate here. Suffice it to say that many philosophers take case judgments from lay people to be relevant. This holds even if it should turn out that most of the classic authors did actually argue for their preferred case judgments.

experiments to test whether reference change occurs if a term—'water' in Experiment 1 and 'salt' in Experiment 2—is consistently applied to a new kind. Whereas the causal source view predicts a significant change in this direction, the causal-historical view does not. The online experiments were implemented on the Unipark platform and participants were recruited from Prolific Academic. As preregistered,[10] each of the two experiments were run until we collected the responses of 200 participants who answered all attention check questions correctly (to reach ~95% power for a difference of 20% in the area of 30% versus 10%). All participants were native English speakers. In Experiment 1, 61% of the respondents indicated themselves to be female and 38% to be male, while the rest identified as non-binary or "other", or preferred not to disclose their gender. Participants were on average 36 years old. In Experiment 2, 60% of the respondents indicated themselves to be female and 39% to be male, while the rest identified as non-binary or "other", or preferred not to disclose their gender. Participants were on average 39 years old. Participants in all experiments were compensated £0.65 for estimated 6 minutes of their time (£6.50/hour; actual median time was 6 minutes 33 seconds).

3.1. Learning Phase

In both experiments, the crucial test questions asked the participants to indicate whether a given statement, formulated in English, is true or false when uttered in a Twin Earth scenario (Putnam 1975) in the 18th century. Pilot studies in which we asked participants to include a short verbal justification of their answers indicated that participants did not consistently interpret our test questions in the way we meant them. Some participants seemed to take the utterer's belief in the proposition expressed by the statement to be sufficient for its truth. Others seemed not to answer whether the statement is true in the context in which it was uttered (in the 18th century on Twin Earth), but rather whether it would be true if uttered here and now (2020 on Earth). As the distinction between truth and appearance (i.e., between mere belief and justified belief) on the one hand, and the distinction between a statement being true when uttered in the 18th century on Twin Earth and it being true here and now, on the other hand, are crucial in order for the experimental data to be valid, we preceded both experiments with

10. You can access the preregistration for Experiment 1 by following this link:
https://osf.io/897us?view_only=0f084d29e2f84d2bb2711c80dead9daa
You can access the preregistration for Experiment 2 by following this link:
https://osf.io/yq3em/?view_only=0bd10b12efac4c4d84d297e3b604613e
The complete material used in Experiment 1 and Experiment 2 can be found here:
https://osf.io/y7kcq/?view_only=7aea21abf75b447abf1faa2ad7b5e305

a learning phase in which these distinctions were explained and practiced. This learning phase involved the two aspects *Truth and falsity* and *Temporal context of the utterance*. Each of these was briefly explained to the participants, followed by two short vignettes and test questions to check whether participants understood the distinctions. This involved the following material:

Truth and falsity

Instruction:

Do not answer whether the person who makes the statement believes it to be true, or has good reason to believe it. Your task is not to judge whether the person in the scenario is lying or justified in believing what she does, but whether the statement in question is *true or false*.

Questions:

At the grocery store, Laura reads a sign saying that the tomatoes are on special offer. However, the special offer stops and the sign is taken away shortly after she reads it, which Laura does not see or realize at all. When she meets her friend Sarah later in the store, she tells her: "The tomatoes are currently on special offer".

When Laura says "The tomatoes are currently on special offer" to her friend Sarah, what she says is literally . . .
[true; false]

Temporal context of the utterance

Instruction:

The language we use is steadily in flux. Some words of English have significantly changed their meanings in the last centuries. Here are two examples:

In the 17th century, the word 'salad' referred only to concoctions of green leaves, whereas now it refers also to cold potato dishes and the like; similarly, the word 'meat' used to refer to any kind of solid food, not just animal flesh.

When you answer whether a statement is true or false, please indicate whether you think the statement is true or false *in the context in which it is uttered*.

Questions:

> In the 17th century, the word 'salad' referred only to concoctions of green leaves, whereas now it refers also to cold potato dishes and the like; similarly, the word 'meat' used to refer to any kind of solid food, not just animal flesh.
>
> When you answer whether a statement is true or false, please indicate whether you think the statement is true or false *in the context in which it is uttered*.
>
> Suppose that, in 1612, the noble man Edward is served a cold dish, consisting of sliced tomatoes and onions, by his servant Peter. Peter says to him: "Sir, this is a fresh salad for you".
>
> When Peter says "Sir, this is a fresh salad for you", what he says is literally . . .
> [true; false]

Participants had three trials to answer all four check-questions correctly. Those who failed to answer them correctly did not proceed to the actual experiment. Later pilot studies confirmed that the learning phase increased the participants' understanding of the test questions significantly.

3.2. Experiment 1

3.2.1. Design, Materials, and Procedure

Experiment 1 tested our hypothesis about reference change due to unintentional (mis)application to another entity, using the example of 'water'. Those participants who mastered the learning phase (roughly 75%) were randomly assigned to one of two conditions, *Rarely* or *Often*, which differed in how often the term 'water' was (mis)applied to a substance composed of XYZ.[11] The first part of the scenario was identical in both conditions and described how a group of people arrive on Twin Earth and explore their surroundings, until one of them, Mary, points to a pond and says:

11. As stated in the preregistrations, we also tested the manipulation in a different design in which the *Truth* question was asked only once (40 years after the arrival on Twin Earth). Since the result patterns were similar (a weaker but significant effect of the frequency manipulation) and given limitations of space, we do not discuss them here. The results and materials for these additional conditions can also be found on the osf.io platform.

"Look! I have found water! There is water in the pond."

This statement was followed by three questions:

> *Truth*$_1$: When Mary says "There is water in the pond", what she says is literally . . .
> [true/false]
> *Reference*$_1$: When Mary uses the term 'water' . . .
> [she refers to a liquid composed of H_2O; she refers to a liquid composed of XYZ; she refers to both: liquids composed of H_2O and liquids composed of XYZ]
> *Meaning*$_1$: In Mary's community in 1750 the term 'water' . . .
> [means «a liquid composed of H_2O»; means «a liquid composed of XYZ»; means «a liquid which is composed of H_2O or XYZ».[12]

After this, participants were presented with the second part of the scenario. In this second part of the scenario, we manipulated how often the term 'water' was (mis)applied to the substance XYZ. The *Often* condition, in which the term was used regularly, reads as follows:

> Mary and the rest of the group finally give up any hope of finding a way back to Earth and decide to settle permanently on Twin Earth.
>
> After the Earthlings have familiarized themselves with their new environment, they start living normal lives on Twin Earth. They build houses in the area where they arrived and they drink the transparent liquid from a small lake nearby their new homes. In addition to drinking it, they also use the transparent liquid in the exact same way that people on Earth use water: they bathe in it, swim in it, use it for cooking and making tea, etc. Just as before, nobody has any clue about the chemical structure of either water or the newly found liquid, so nobody can tell that there is any difference between them. They all believe that the liquid they found on Twin Earth is water. Moreover, they keep using the word 'water' to refer to this liquid.
>
> 40 years later, referring to the liquid on Twin Earth as 'water' has become a common practice in Mary's group. One day, Mary and her fellow Earthlings, who now have spent most of their life on Twin Earth, go searching for another place to quench their thirst. After a while, they

12. After these three questions, participants were asked to answer two attention checks (about the chemical composition of the liquid substance on Twin Earth, and whether Mary and her group know its composition).

> arrive at exactly the same pond that Mary and her fellows found right after being teleported to Twin Earth. Pointing to the pond, Mary says: "Look! I have found water! There is water in the pond."

In the *Rarely* condition, in which the term was virtually never used, the scenario continues as follows:

> Due to a side-effect of teleportation, Mary and the rest of the group suddenly become very tired. They fall into an unusually long and extremely deep sleep, and do not awake for the next 40 years.
>
> While asleep, time passes at its normal speed: the tides go in and out, the trees grow taller, predators hunt their prey, and birds sing their songs. Of course, neither Mary nor any other member of the group is aware of any of these things. Nonetheless, the group members grow old just as they would if they had been awake all the time.
>
> 40 years later, Mary and her fellows finally wake up. Even though they have spent the last 40 years on Twin Earth, they barely had a chance to actively live there, due to their unusually long sleep. They still know very little about Twin Earth, and they had almost no opportunity to talk about their experiences after their arrival. When they wake up again, they find themselves exactly where Mary noticed the pond 40 years earlier. A bit perplexed about what happened to them, Mary remarks: "Look! I have found water! There is water in the pond."

Afterwards, participants were again asked the three questions about *Truth*, *Reference*, and *Meaning*.[13]

> *Truth*$_2$: When Mary now says "There is water in the pond", what she says is literally . . .
> [true; false]
> *Reference*$_2$: When Mary uses the term 'water' now . . .
> [she refers to a liquid composed of H_2O; she refers to a liquid composed of XYZ; she refers to both: liquids composed of H_2O and liquids composed of XYZ]
> *Meaning*$_2$: In Mary's community in 1790 (40 years after they landed on Twin Earth) the term 'water' . . .
> [means «a liquid composed of H_2O»; means «a liquid composed of XYZ»; means «a liquid which is composed of H_2O or XYZ»]

13. After answering the questions about *Truth*, *Reference*, and *Meaning*, participants were again asked two attention check questions on a different screen.

The crucial (preregistered) measure was about the truth of Mary's statement at the two points of time. The (preregistered) operationalization of reference change was judging Mary's statement to be false when first assessed (when the group arrives on Twin Earth, $Truth_1$) but true at the later time (40 years later, $Truth_2$). Our (preregistered) prediction was that the proportion of reference change in *Often* would be significantly higher than in *Rarely*. We included the questions about *Reference* and *Meaning* to see (exploratorily) whether they follow the same pattern as responses for the *Truth* question.

3.2.2. Results and Discussion

The results of the experiment are shown in Figure 1 and Figure 2. Regarding the crucial *Truth* question, only a minority considered Mary's statement to be true at the time of arrival (only 13% in *Rarely* and 16% in *Often*). In the *Rarely* condition, the percentage of participants considering the statement to be true when Mary made it 40 years later did not change significantly (11%). In the *Often* condition, by contrast, the percentage of participants judging the statement to be true increased to 40%.

Similarly, the proportions of reference change, that is, judging Mary's statement to be false on arrival on Twin Earth but true 40 years later, was significantly higher in the *Often* condition (32%), as compared to the *Rarely* condition (2%), $z=5.65$, $p < .001$. Hence, the finding confirmed our prediction that reference change will occur due to constant (mis)application of a term. This claim gains further support from the results for the *Reference* and *Meaning* questions

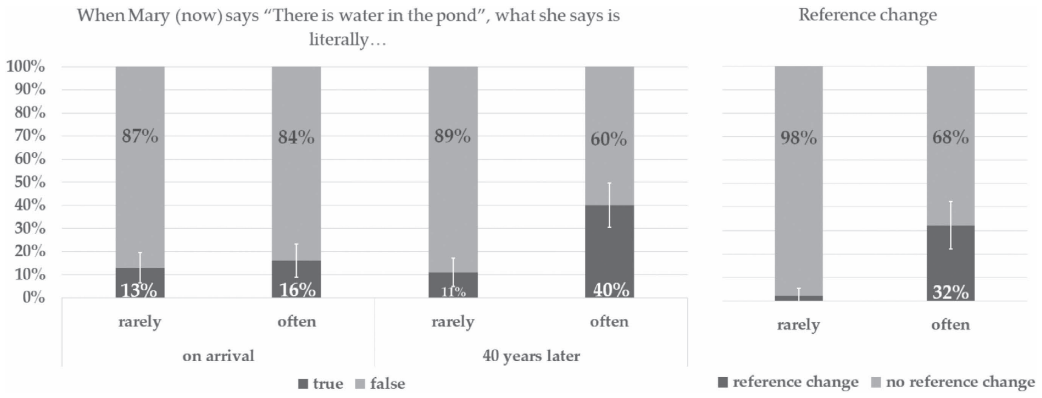

Figure 1. Left four bars: Percentages of participants choosing the "true" or "false"-option as a function of when the truth of the statement "There is water in the pond" was assessed and of how often the term 'water' was (mis)applied to XYZ. Right two bars: Percentages of participants indicating reference change for the term 'water', i.e., judging the statement to be false on arrival but true forty years later, again as a function of how often the term 'water' was used. Error bars indicate 95% CI.

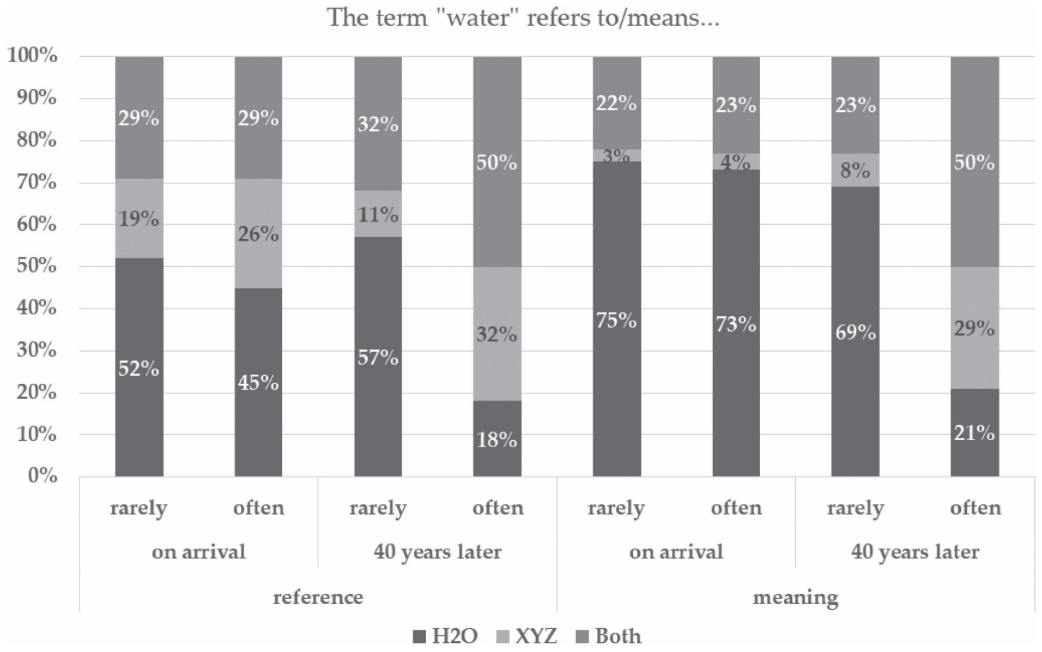

Figure 2. Percentages of participants choosing the "H_2O", "XYZ", or "Both"-option as a function of the questions asked (reference on the four leftmost bars and meaning on the four rightmost bars), when the question was asked (at the time of arrival and 40 years later), and how often the term 'water' was (mis)applied (rarely versus often).

(see Figure 2). When the term 'water' was constantly (mis)applied, the proportions of participants judging the term to refer to or mean H_2O decreased, while the proportions of participants judging the term to now refer to or mean XYZ increased.

3.3. Experiment 2

3.3.1. Design, Materials, and Procedure

Experiment 2 was structurally analogous to Experiment 1 but involved the term 'salt' instead of 'water'. Again, those participants who mastered the learning phase (roughly 75%) were randomly assigned to one of two conditions, *Rarely* or *Often*, which differed in how often the term 'salt' was (mis)applied to a substance composed of XYZ. The first part of the scenario was identical in both conditions and described how a group of people arrive on Twin Earth and find a cave containing a white crystalline substance, whereupon one of them, Peter, says:

> "Look! I have found salt! There is salt in this cave."

This statement was followed by three questions:

> *Truth*$_1$: When Peter says "There is salt in this cave", what he says is literally . . .
> [true; false]
> *Reference*$_1$: When Peter uses the term 'salt' . . .
> [he refers to a substance composed of NaCl; he refers to a substance composed of XYZ; he refers to both: a substance composed of NaCl and a substance composed of XYZ]
> *Meaning*$_1$*:* In Peter's community in 1750 the term 'salt' . . .
> [means «a substance composed of NaCl»; means «a substance composed of XYZ»; means «a substance which is composed of NaCl or XYZ»[14]

In the second part of the scenario, we manipulated how often the term 'salt' was (mis)applied to the substance XYZ. The *Often* condition, in which the term was used regularly, reads as follows:

> Peter and the rest of the group finally give up any hope of finding a way back to Earth and decide to settle permanently on Twin Earth.
>
> After the Earthlings have familiarized themselves with their new environment, they start living normal lives on Twin Earth. They build houses in the area where they arrived and they frequently visit a cave which they found nearby their new home to get some of the white crystal substance. They use this substance to make their food tastier, to treat stains on their clothes, or to prevent their pathways from freezing. Just as before, nobody has any clue about the chemical structure of either salt or the newly found substance, so nobody can tell that there is any difference between them. They all believe that the white crystal substance they found on Twin Earth is salt. Moreover, they keep using the word 'salt' to refer to this substance.
>
> 40 years later, referring to the white crystal substance on Twin Earth as 'salt' has become a common practice in Peter's group. One day, Peter and his fellow Earthlings, who now have spent most of their life on Twin Earth, go searching for another cave to collect more of the white crystal substance. After a while, they arrive at exactly the same cave that Peter and his fellows found right after being teleported to Twin Earth. Entering the cave, Peter says: "Look! I have found salt! There is salt in this cave."

14. Again, after these three questions, participants were asked to answer two attention checks (about the chemical composition of the white crystalline substance on Twin Earth, and whether Peter and his group know its composition).

In the *Rarely* condition, in which the term was virtually never used, the scenario continues as follows:

> Due to a side-effect of teleportation, Peter and the rest of the group suddenly become very tired. They fall into an unusually long and extremely deep sleep, and do not wake up for the next 40 years.
>
> While asleep, time passes at its normal speed: the tides go in and out, the trees grow taller, predators hunt their prey, and birds sing their songs. Of course, neither Peter nor any other member of the group is aware of any of these things. Nonetheless, the group members grow old just as they would if they had been awake all the time.
>
> 40 years later, Peter and his fellows finally wake up. Even though they have spent the last 40 years on Twin Earth, they barely had a chance to actively live there, due to their unusually long sleep. They still know very little about Twin Earth, and they had no opportunity to talk about their experiences after their arrival. When they wake up again, they find themselves exactly where Peter noticed the cave 40 years earlier. A bit perplexed about what happened to them, Peter remarks: "Look! I have found salt! There is salt in the cave."

Afterwards, participants were again asked the three questions about *Truth*, *Reference*, and *Meaning*.[15]

> *Truth*$_2$: When Peter now says "There is salt in the cave", what he says is literally . . .
> [true; false]
> *Reference*$_2$: When Peter uses the term 'salt' now . . .
> [he refers to a substance composed of NaCl; he refers to a substance composed of XYZ; he refers to both: a substance composed of NaCl and a substance composed of XYZ]
> *Meaning*$_2$: In 1790 (40 years after they landed on Twin Earth), in Peter's community on Twin Earth the term 'salt' . . .
> [means «a substance composed of NaCl»; means «a substance composed of XYZ»; means «a substance which is composed of NaCl or XYZ»

Again, the crucial (preregistered) measure was about the truth of Peter's statement at the two points of time, and the (preregistered) operationalization of reference change was judging Peter's statement to be false when first assessed

15. After answering the questions about *Truth*, *Reference*, and *Meaning*, participants were again asked two attention check questions on a different screen.

(when the group arrives on Twin Earth) but true at the later time (40 years later). Our (preregistered) prediction was that the proportion of reference change in *Often* would be significantly higher than in *Rarely*. We included the questions about *Reference* and *Meaning* to see (exploratorily) whether they follow the same pattern as responses for the *Truth* question.

3.3.2. Results and Discussion

The results of the experiment are shown in Figure 3 and Figure 4.

Regarding the *Truth* question, few considered Peter's statement to be true at the time of arrival (only 12% in *Rarely* and 19% in *Often*). In the *Rarely* condition, this did not change significantly (16%) when the statement was assessed for the second time. By contrast, in the *Often* condition the percentage of participants judging the statement to be true increased to 53%. The proportions of reference change, that is, judging Peter's statement to be false on arrival on Twin Earth but true 40 years later, was significantly higher in the *Often* condition (47%) than in the *Rarely* condition (9%), $z=5.52$, $p < .001$. Hence this finding, too, confirmed our prediction that reference change will occur due to constant (mis)application of a term. The results for the *Reference* and *Meaning* questions (see Figure 4) further support this claim. When the term 'salt' was constantly (mis)applied, the proportions of participants judging that the term refers to or means NaCl decreased,

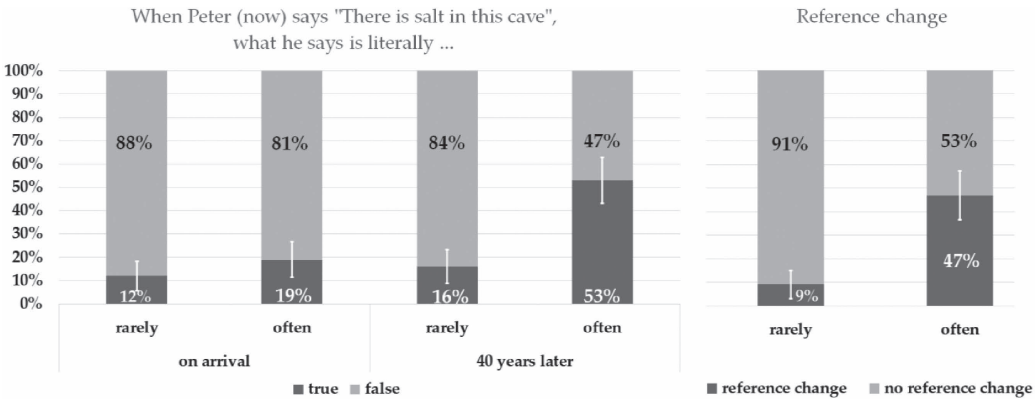

Figure 3. Left four bars: Percentages of participants choosing the "true" or "false"-option as a function of when the truth of the statement "There is salt in this cave" was assessed and of how often the term 'salt' was (mis)applied to XYZ. Right two bars: Percentages of participants indicating reference change for the term 'salt', i.e., judging the statement to be false on arrival but true forty years later, again as a function of how often the term 'salt' was used. Error bars indicate 95% CI.

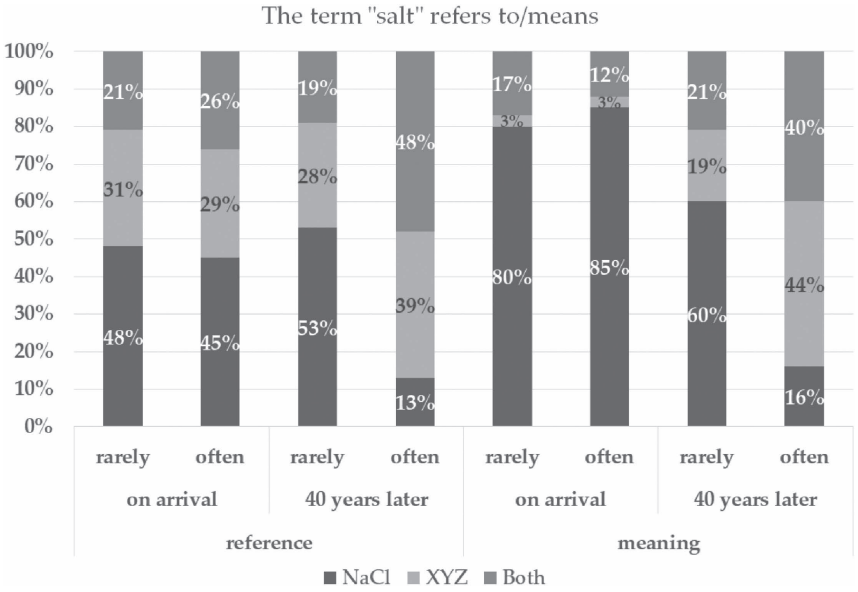

Figure 4. Percentages of participants choosing the "NaCl", "XYZ", or "Both"-option as a function of the questions asked (*reference* on the four leftmost bars and *meaning* on the four rightmost bars), when the question was asked (at the time of arrival and 40 years later), and how often the term 'salt' was (mis)applied (rarely versus often).

while the proportions of participants judging that the term now refers to or means XYZ increased.

4. General Discussion

4.1. Methodological Discussion

Before we turn to interpreting our results and their philosophical significance, a few words in defense of our methodological approach are in order. Recent debates about experimental semantics reveal some controversy about how to elicit valid responses to questionnaires used in experimental semantics (e.g., Devitt 2011; 2012; Devitt & Porot 2018; Machery, Olivola, & de Blanc 2009; Martí 2009; 2012; 2014; 2020). Martí argues in a series of papers (2009; 2012; 2014; 2020) that questions about reference and meaning, even in covert forms such as "What is John talking about?" (Machery et al. 2004) target metalinguistic rather than linguistic intuitions, which she deems less relevant for theorizing about reference. The correct theory of reference, she argues, should be in line with how lay people actually *use*, for example, names and natural kind terms, but not necessarily

with how they *think* that reference works in the case of names and natural kind terms. She therefore claims that a better way to test the adequacy of theories of reference is to provide participants with elicited production tasks, that is, questions with open answer fields which prompt people to use the relevant term in one way or another.

Does Martí's criticism undermine our experimental setup? We have two reasons for thinking that this is not so. First, we do not think that our primary test questions about truth and falsity elicit metalinguistic intuitions about how reference works rather than directly linguistic intuitions about what the terms in question refer to. In fact, this is the reason why we used truth value judgments as our primary (preregistered) measure for reference change and placed our exploratory questions about reference and meaning on a separate page, thus making sure that these questions would not affect the participants' answers. It is moreover difficult to see how truth value judgments should *not* be relevant for theorizing about reference. After all, these judgments indicate what native speakers take to be the truth conditions of sentences containing the relevant terms. In a controlled setting like ours, this gives us a good measure of the terms' reference. This has recently been recognized even by the early critics of experimental semantics (Devitt & Porot 2018; Martí 2020). Secondly, the elicited production tasks suggested by Martí come with many problems of their own. As Machery et al. (2009) mention, responses to open answer questions are notoriously hard to interpret. Moreover, given that Martí proposes to present participants with questions seemingly about issues other than language (e.g., whether Gödel in Kripke's scenario is blameworthy), there is the risk that convoluting factors obscure the data. All in all, we take the collection of truth value judgments to be a better guide to people's intuitions about matters of reference.

4.2. Discussion of the Results

It is now time to discuss the implications of our results. The main part of the discussion will be devoted to the question of which view is ultimately better supported by our data: Kripke's causal-historical theory of reference, or rather the causal source view as advocated by Evans and Devitt. Before we get to this issue, however, two other striking features of our data deserve discussion.

A first noteworthy result of our experiments is that a large majority of subjects seem to have externalist intuitions about the reference of natural kind terms. In both experiments, only ~15% of the participants took the relevant statement, that is, "This is water" or "This is salt", uttered by someone pointing to a substance with a chemical structure different from water or salt, to be true when assessed at t_1. This means that ~85% of the participants took the truth of these statements

to be partly determined by the (unknown) underlying chemical structure of the respective natural kind. Our results thus confirm earlier findings by Jylkkä, Railo, and Haukioja (2009), but are in tension with the findings of Genone and Lombrozo (2012), who report much lower numbers for the externalist judgments. We suspect that the learning phase which preceded our experiments, but which was lacking in other studies, plays an important role in explaining the differences between their findings and ours. Given how crucial the learning phase turned out to be for an adequate understanding of the issue, the lower ratings found in earlier studies could be partly due to misunderstandings on the part of the participants.

Another noteworthy result is the striking difference in responses to our truth and reference questions in both experiments, especially when assessed at t_1.[16] Whereas for our truth questions, a large majority of the participants in both experiments indicated the statement "This is water" or "This is salt" to be false when uttered upon arrival (between 81% and 88%, depending on the condition), almost half of the participants (between 48% and 55%) judged 'water' and 'salt' to refer either to XYZ or to [H_2O or XYZ] upon arrival. This implies that many participants (between 26% and 39%) took the initial judgments, for example, "This is water", to be false *while at the same time* thinking that the term in question refers to the very entity that the protagonists of the vignettes think they are talking about, for example, a substance composed of XYZ. As philosophers of language, we are used to the idea that facts about reference determine truth conditions, and hence that judgments about reference should be aligned with judgments about truth. So why is there such a glaring divergence between the responses we have received for our truth and reference questions?

The most plausible hypothesis is that this difference is again due to the learning phase that preceded both experiments. This learning phase was centered on truth and falsity rather than meaning or reference. As mentioned above, earlier pilot studies showed that the learning phase played a crucial role in making participants understand the test questions correctly. In particular, the learning phase taught the participants that we are asking for an objective assessment rather than a subjective one, that is, whether the statement is actually true rather than just believed to be true by the speaker. Since the participants did not undergo any such training about reference, some of them might have answered the relevant questions in terms of (something like) speaker's reference (cf. Kripke 1977) rather than semantic reference: they might have answered what the speaker in the scenario intends to refer to, rather than what the term semantically refers to. Considering that the protagonists of our vignettes clearly intend to refer to the substances they encounter on Twin Earth, this would explain why the ratings

16. Thanks to an anonymous reviewer for raising this issue.

for the relevant answer options are so much higher than the corresponding truth ratings.

Let us now turn to the question which of the two views, the causal-historical view or the causal source view, is better supported by our data. The first thing to note here is that not all of our data actually speak to this question. In particular, roughly 15% (between 12% and 18%, depending on the condition) answered that the statement "This is water" or "This is salt" is already true when assessed at t_1. Since neither the causal-historical view nor the causal source view are compatible with such ratings, we will neglect them in what follows. Suffice it to say that, as far as the assessment at t_1 is concerned, both views seem to be well supported by the data.

The crucial measure for deciding between the two views is how many of those who took the relevant statement to be false at t_1 changed their opinion when they were asked to make the same assessment at t_2. We found that 32% of the responses we received for Experiment 1 and 47% of those received for Experiment 2 exhibit this pattern. Moreover, our exploratory test questions about reference and meaning display similar patterns. In both experiments, the reference ratings for H_2O/NaCl dropped from 45% at t_1 to a mere ~14% at t_2. In Experiment 1, the meaning ratings for H_2O dropped from 73% at t_1 to 21% at t_2; in Experiment 2 they dropped even further, from 85% to 16%.

These results indicate that a large number of ordinary native English speakers think that terms like 'water' or 'salt' switch reference (or change their meanings) after being consistently (mis)applied to new kinds of entities over long periods of time. This holds despite the fact that in the tested vignettes nobody had the intention of using these terms differently from before. Since Kripke's causal-historical view predicts that there should be no reference change as long as the terms are properly introduced and nobody has the intention of changing their reference, responses that nevertheless indicate reference change can be seen as empirical evidence against this view. Moreover, since causal source views in the spirit of Evans and Devitt equally capture the typical externalist case judgment (e.g., in the Gödel-Schmidt case or in the classic Twin Earth case), but also predict a significant proportion of reference change in cases of the above type, our data provide empirical evidence in favor of these views.

Nonetheless, the interpretation of our data demands caution. As illustrated above, our experiments clearly indicate reference change across all question-types (truth, meaning, reference). For some questions, these ratings even constitute a majority. For most questions, however, and in particular for our crucial truth questions, they did not. In Experiment 1, 68% of those participants who took "This is water" to be false at t_1 still thought the same at t_2. In Experiment 2, the effect was stronger, but still 53% indicated that there was no reference change. How are we to interpret these responses? Shouldn't they be interpreted

as counting in favor of the causal-historical view and against the causal source view? If so, then doesn't the entirety of our data support these two views (at best) about equally well?[17]

We think that drawing this conclusion would be too quick. It is true that the results are mixed and that we cannot claim a clear majority in favor of reference change through all conditions. Thus, if the causal source view were committed to the claim that the reference of terms like 'water' or 'salt' unambiguously refers to a different kind of entity after being consistently (mis)applied for 40 years, then our results would be at least as problematic for the causal source view as they are for the causal-historical view. However, our discussion of the causal source view in Section 2, and in particular how it accounts for reference change, clearly shows that this view is not so committed. Recall that, according to the causal source view, for a term to switch reference from one (kind of) entity to another, all—or at least most—traces of information that are causally derived from the former have to be erased from the linguistic community. Since people exchange information with each other all the time, such traces of information are usually kept alive within a community for a very long time.

In our test vignettes, people live on Twin Earth for 40 years, and interact with substances that they take to be water and salt respectively. After 40 years, a lot of the information they associate with 'water' and 'salt' will be derived from their experiences on Twin Earth. But the protagonists of our vignettes as well as their respective communities were born and raised on Earth, not on Twin Earth. Even after 40 years of living on Twin Earth, they can plausibly be imagined as having vivid memories of their water- and salt-involving experiences on Earth. It is only to be expected that some traces of information that are derived from water and salt should still remain in their communities. In light of this, the demand for a strong majority in favor of reference change in our experiments does not do justice to the details of the causal source view.

Nonetheless, as they stand, our data do not allow us to claim victory for the causal source view either. This is because we do not know enough about the attitudes of those participants who judged "This is water" or "This is salt" to remain false when assessed at t_2. According to the above hypothesis, at least some of these participants will eventually come to agree that the statement is true; namely when more of the information traces that are causally derived from water fade from the imagined linguistic community. But this is admittedly quite speculative, since it could also be that all of these participants are die-hard Kripkeans and will never agree with the truth of the statement, come what may.

17. Thanks to an anonymous reviewer and the area editor of *Ergo* for prompting a more detailed discussion of this issue.

To test how many of those participants whose answers seem incompatible with the causal source view agree that reference change will occur later on, we have conducted an (exploratory) follow-up experiment. The main part of this experiment was identical to the *Often* condition of Experiment 2 above. It was run until 100 participants indicated at t_2 that no reference change has occurred. These participants were directed to another page and asked:

> You have answered that when Peter now, after 40 years of living on Twin Earth, says 'This is salt', what he says is literally false.
>
> Please indicate which of the following two options describes your opinion better:
>
> (a) Peter's utterance 'This is salt' is false, and no matter for how long people keep living on Twin Earth in the described way, when somebody points to XYZ and calls it 'salt', this utterance will always be false.
> (b) Peter's utterance 'This is salt' is false, but if people keep living on Twin Earth in the described way for much longer, then eventually, when somebody points to XYZ and calls it 'salt', this utterance will be true.

In this experiment, 45% of all the participants indicated reference change, that is, they took the statement to be false upon arrival but true 40 years later. Interestingly, however, 28% of those who judged the statement to be false at both t_1 and at t_2 still chose answer (b) above. Since answer (b) is clearly incompatible with the causal-historical view, but in line with the causal source view, these 28% also speak against the former and in favor of the latter. If we do the math, this leaves us with 61% of the answers being incompatible with the causal-historical view and in line with the causal source view, versus 39% being in line with the causal-historical view, and incompatible with the causal source view. This amounts to a narrow majority in favor of the causal source view.

Moreover, from the perspective of the causal source view, one might wonder whether even the remaining 39% are staunch Kripkeans. As mentioned in Section 2, reference change according to the causal source view is rather complex. Anything that has an effect on the proportion of information causally derived from *y* instead of *x* that is in circulation within a linguistic community plays a role therein. This involves contingent demographic facts, facts about who gets to read what book or talks to which person, the amount of knowledge and beliefs that people have prior to the exchange of the two (kinds of) objects, which important discoveries they make afterwards, how much they interact with each other and with the new (kind of) object, and potentially much more. Obviously, the story given in our vignettes did not and could not include all these potentially

relevant factors. It is therefore at least possible that different or more complex vignettes would yield stronger results in favor of the causal source view. Alas, a deeper investigation into the levers of reference change and how they interact with each other will have to wait for another occasion. For now, we conclude that our experiments provide interestingly new, though certainly not decisive, evidence in favor of the causal source view—a view that has received far less attention than it deserves.

Acknowledgments

The materials of this paper were presented at the EXTRA.1 workshop (Ruhr University Bochum) and at the Experimental Philosophy of Language and Metaethics Workshop (Ruhr University Bochum). We thank the audiences for their helpful feedback. For very helpful comments on previous versions of the paper, we would like to thank Michael Devitt, Nat Hansen, John Horden, Edouard Machery, Genoveva Martí, Sigurd Jorem, Guido Löhr and two anonymous reviewers as well as the area editor of *Ergo*. Steffen Koch and Alex Wiegmann's work on this paper was supported by an Emmy Noether grant of the *German Research Foundation* (DFG), project number 391304769.

References

Adleberg, Toni, Morgan Thompson, and Eddy Nahmias (2015). Do Women and Men Have Different Philosophical Intuitions? Further Data. *Philosophical Psychology*, *28*(5), 615–41.

Beebe, James R. and Ryan J. Undercoffer (2015). Moral Valence and Semantic Intuitions. *Erkenntnis*, *80*(2), 445–66.

Beebe, James R. and Ryan J. Undercoffer (2016). Individual and Cross-Cultural Differences in Semantic Intuitions: New Experimental Findings. *Journal of Cognition and Culture*, *16*(3–4), 322–57.

Braisby, Nick, Bradley Franks, and James Hampton (1996). Essentialism, Word Use, and Concepts. *Cognition*, *59*(3), 247–74.

Buckwalter, Wesley and Stephen Stich (2015). Gender and Philosophical Intuition. In Joshua Knobe and Shaun Nichols (Eds.), *Experimental Philosophy* (Vol. 2, 307–46). Oxford University Press.

Cappelen, Herman (2012). *Philosophy without Intuitions*. Oxford University Press.

Devitt, Michael (1981). *Designation*. Columbia University Press.

Devitt, Michael (2011). Experimental Semantics. *Philosophy and Phenomenological Research, 82*(2), 418–35.

Devitt, Michael (2012). Whither Experimental Semantics? *THEORIA. An International Journal for Theory, History and Foundations of Science, 27*(1), 5–36.

Deutsch, Michael (2015). *The Myth of the Intuitive*. MIT Press.

Devitt, Michael and Nicolas Porot (2018). The Reference of Proper Names: Testing Usage and Intuitions. *Cognitive Science*, *42*(5), 1552–85.

Devitt, Michael and Kim Sterelny (1999). *Language and Reality*: *An Introduction to the Philosophy of Language* (2nd ed.). Blackwell.

Dickie, Imogen (2015). *Fixing Reference*. Oxford University Press.

Evans, Gareth (1973). The Causal Theory of Names. *Aristotelian Society Supplementary Volume*, *47*(11), 187–225.

Evans, Gareth (1982). *The Varieties of Reference*. Oxford University Press.

Fine, Arthur (1975). How to Compare Theories: Reference and Change. *Noûs*, *9*(1), 17–32.

Frege, Gottlob (1948). Sense and Reference. *The Philosophical Review*, 57(3), 209–30.

Genone, James and Tania Lombrozo (2012). Concept Possession, Experimental Semantics, and Hybrid Theories of Reference. *Philosophical Psychology*, *25*(5), 717–42.

Häggqvist, Sören and Åsa Wikforss (2015). Experimental Semantics: The Case of Natural Kind Terms. In Jussi Haukioja (Ed.), *Advances in Experimental Philosophy of Language* (109–38). Bloomsbury Publishing.

Horvath, Joachim and Steffen Koch (2021). Experimental Philosophy and the Method of Cases. *Philosophy Compass, 16*(1):e12716.

Jylkkä, Jussi, Henry Railo, and Jussi Haukioja (2009). Psychological Essentialism and Semantic Externalism: Evidence for Externalism in Lay Speakers' Language Use. *Philosophical Psychology, 22*(1), 37–60.

Koch, Steffen (2021a). The Externalist Challenge to Conceptual Engineering. *Synthese, 198*(1), 327–348. https://doi.org/10.1007/s11229-018-02007-6

Koch, Steffen (2021b). Engineering What? On Concepts in Conceptual Engineering. *Synthese, 199*(1–2), 1955–1975. https://doi.org/10.1007/s11229-020-02868-w

Koch, Steffen (2021c). There Is No Dilemma for Conceptual Engineering. Reply to Max Deutsch. *Philosophical Studies, 178*(7), 2279–2291. https://doi.org/10.1007/s11098-020-01546-4

Koslicki, Kathrin (2008). Natural Kinds and Natural Kind Terms. *Philosophy Compass*, *3*(4), 789–802.

Kripke, Saul A. (1977). Speaker's Reference and Semantic Reference. In Peter A. French, Theodore Edward Uehling, and Howard K. Wettstein (Eds.), *Studies in the Philosophy of Language* (255–76). Morris.

Kripke, Saul A. (1980). *Naming and Necessity*. Harvard University Press.

LaPorte, Joseph (2004). *Natural Kinds and Conceptual Change*. Cambridge University Press.

Machery, Edouard (2017). *Philosophy within Its Proper Bounds*. Oxford University Press.

Machery, Edouard, Max Deutsch, Justin Sytsma, Ron Mallon, Shaun Nichols, and Stephen Stich (2004). Semantics, Cross-Cultural Style. *Cognition*, *92*(3), 1–12.

Machery, Edouard, Max Deutsch, Justin Sytsma, Ron Mallon, Shaun Nichols, and Stephen Stich (2010). Semantic Intuitions: Reply to Lam. *Cognition*, *117*(3), 361–66.

Machery, Edouard, Christopher Y. Olivola, and Molly de Blanc (2009). Linguistic and Metalinguistic Intuitions in the Philosophy of Language. *Analysis*, *69*(4), 689–94.

Machery, Edouard, Justin Sytsma, and Max Deutsch (2015). Speaker's Reference and Cross-Cultural Semantics. In Andrea Bianchi (Ed.), *On Reference* (62–76). Oxford University Press.

Martí, Genoveva (2009). Against Semantic Multi-Culturalism. *Analysis, 69*(1), 42–48.

Martí, Genoveva (2012). Empirical Data and the Theory of Reference. In William P. Kabasenche, Michael O'Rourke, and Matthew H. Slater (Eds.), *Reference and Referring* (63–82). MIT Press.

Martí, Genoveva (2014). Reference and Experimental Semantics. In Edouard Machery (Ed.), *Current Controversies in Experimental Philosophy* (17–26). Routledge.

Martí, Genoveva (2020). Experimental Semantics, Descriptivism and Anti-Descriptivism. Should We Endorse Referential Pluralism? In Andrea Bianchi (Ed.), *Language and Reality from a Naturalistic Perspective: Themes from Michael Devitt* (329–44). Springer.

Putnam, Hilary (1975). The Meaning of 'Meaning'. In *Philosophical Papers* (Vol. 2, Mind, Language and Reality, 215–71). Cambridge University Press.

Russell, Bertrand (1905). On Denoting. *Mind*, *14*(56), 479–93.

Searle, John R. (1958). Proper Names. *Mind*, *67*(266), 166–73.

Seyedsayamdost, Hamid (2015). On Gender and Philosophical Intuition: Failure of Replication and Other Negative Results. *Philosophical Psychology*, *28*(5), 642–73.

Sytsma, Justin, Jonathan Livengood, Ryoji Sato, and Mineki Oguchi (2015). Reference in the Land of the Rising Sun: A Cross-Cultural Study on the Reference of Proper Names. *Review of Philosophy and Psychology*, *6*(2), 213–30.

Thomasson, Amie L. (2007). *Ordinary Objects*. Oxford University Press.

Tobia, Kevin P, George E. Newman, and Joshua Knobe (2020). Water Is and Is Not H_2O. *Mind and Language, 35*(2), 183–208.