Received: 30 April 2018 | Revised: 6 September 2018 | Accepted: 13 November 2018

DOI: 10.1111/mila.12234

ORIGINAL ARTICLE

WILEY

Water is and is not H_2O

Kevin P. Tobia[1] | **George E. Newman**[2] | **Joshua Knobe**[3]

[1]Department of Philosophy, Yale University, New Haven, Connecticut

[2]School of Management, Yale University, New Haven, Connecticut

[3]Program in Cognitive Science and Department of Philosophy, Yale University, New Haven, Connecticut

Correspondence
Kevin P. Tobia, Department of Philosophy, Yale University, New Haven, CT 06511.
Email: kevin.tobia@yale.edu

The Twin Earth thought experiment invites us to consider a liquid that has all of the superficial properties associated with water (clear, potable, etc.) but has entirely different deeper causal properties (composed of "XYZ" rather than of H_2O). Debates about natural kind concepts have sought to accommodate an apparent fact about ordinary people's judgments: Intuitively, the Twin Earth liquid is not water. We present results showing that people do not have this intuition. Instead, people tend to judge that there is a sense in which the liquid is not water but also a sense in which it *is* water.

KEYWORDS
concepts, dual character, experimental philosophy, natural kinds, Twin Earth

1 | INTRODUCTION

Imagine a liquid that is exactly like H_2O in its *superficial properties.* It has the same color, texture, and taste of H_2O and it quenches thirst in just the same way. However, this liquid is completely different when it comes to its *deeper causal properties.* When chemists examine it, they find that it is not composed of H_2O but of some very different chemical compound. Is this liquid water?

This is the well-known "Twin Earth" thought experiment. Since this thought experiment was first introduced, the usual view within philosophy has been *no*, the Twin Earth liquid is not water.

Although this thought experiment was originally introduced to illuminate questions in the theory of reference, it has also played a crucial role in empirically informed debates about natural kind concepts (e.g., Machery & Seppälä, 2011; Margolis, 1998; Nichols, Pinillos & Mallon, 2016; Strevens, 2000). Within this latter stream of research, facts about patterns in people's ordinary judgments serve as an important source of evidence. If we assume that people's ordinary judgment in the Twin Earth case is *no*, we thereby acquire evidence for a view according to which people regard deeper causal properties as the key criteria for membership in natural kind categories. On such a view, when it comes to a natural kind like water, it is not the superficial properties (e.g., clear, potable, etc.) but the deeper causal properties (e.g., H_2O) that are seen as criterial for category membership.

A series of classic empirical studies seemed to provide evidence for the view that deeper causal properties play a key role in judgments of natural kind category membership (e.g., Gelman, 2003; Keil, 1989). Surprisingly, however, more recent studies call into question the view that ordinary people share the assumed Twin Earth judgment (Corcoran, 2018; Miller & Nahmias, 2018). These studies indicate that people actually *do not* strongly endorse the claim that the Twin Earth liquid is not water. This finding might lead one to the radical conclusion that there was never anything right in the assumption that people apply natural kind terms using criteria that involve only deeper causal properties.

We present four studies that show that something is true on both sides. When thinking about Twin Earth, people do not flatly endorse the standard philosophical intuition, nor do they firmly reject it. Instead, they assent to two distinct claims:

1. There is a sense in which the liquid is water.
2. Ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid is not truly water at all.

We call this pattern of judgment a "dual character" pattern. It provides evidence that people's intuitions about natural kinds are driven by two distinct sets of criteria. The complex and ambivalent reaction people have to Twin Earth cases arises from a conflict between these criteria.

We suggest that these results support views that posit two distinct sets of criteria for natural kind category membership. In other words, our findings provide evidence against views that posit just *one* criterion involving deeper causal properties (e.g., H_2O), or *one* criterion involving a set of superficial properties (e.g., clear, potable, etc.), or even *one* set of criteria involving both kinds of properties (e.g., H_2O, clear, potable, etc.). Instead, our results provide evidence in favor of views that posit *two* distinct sets of criteria (one set based on deeper causal properties, another based on more superficial, observable properties). When those criteria deliver different verdicts (e.g., in the Twin Earth case), people make a more complex judgment: The Twin Earth liquid is water (in one sense) *and* is not water (in another sense).

We conclude by considering some possible implications of the results for theories of reference. Of course, facts about people's ordinary intuitions are just one source of evidence in these inquiries, and further work might shed additional light on these questions. However, insofar as empirical and philosophical theories draw on evidence from people's ordinary intuitions, we contend that they should focus not on a mistaken assumption about people's judgments but instead on this dual character intuition.

2 | TWIN EARTH INTUITIONS AND THE PSYCHOLOGY OF CONCEPTS

First, consider theories in the philosophy of psychology about natural kind concepts. Different theories make different predictions about people's Twin Earth intuitions. Thus, an investigation of Twin Earth intuitions can provide evidence that helps to decide between these theories.

Before turning to the different theories and their empirical predictions, we should note one broad assumption shared by all of the theories we will be considering. Research across philosophy and psychology suggests that natural kind concepts are associated with both (a) representations of easily observable s*uperficial properties*, and (b) representations of less easily observable *deeper causal properties.* For example, people associate the concept water with a representation of certain

superficial properties (colorless, tasteless, drinkable, etc.) and also a representation of certain deeper causal properties (e.g., H_2O; see generally Cimpian & Markman, 2009; Gelman, 2003; Gelman & Markman, 1987; Gelman & Wellman, 1991; Kalish & Gelman, 1992; Newman & Keil, 2008; Opfer & Siegler, 2004). Different researchers may adopt radically different views about the nature of concepts or about how people make categorization judgments, but at the very least, everyone should agree that people do have these representations.

One common view is that these two representations have two different formats. Specifically, some research suggests that superficial properties are represented using *prototypes* (e.g., Hampton, 1998; Medin & Ortony, 1989), while other research suggests that the deeper causal properties might be represented using *theories* (Carey, 1985; Keil, 1989; Keil & Wilson, 2000; Murphy & Medin, 1985). There is an impressive body of existing work on these two kinds of representations and their distinctive roles in human cognition.

Our aim in this paper is not to further explore questions about the nature of these representations, and we remain broadly agnostic about how people represent superficial and deep properties. The question we ask is about the categorization process for natural kind concepts. We will be considering a range of possible types of theories. Specifically, categorization judgments might be based on:

1. Just an entity's superficial properties.
2. Just an entity's causal properties.
3. Some mixture of both together.
4. The entity's causal properties and the superficial properties, each separately.

Each of these types of theories makes a different prediction, namely: The Twin Earth liquid should be judged as (a) simply water; (b) simply not water; (c) water, not water, or sort-of water; or (d) in one sense water, but in another sense not water at all. In other words, discovery of certain patterns of judgment would be inconsistent with some theories' predictions. For example, judgment that the Twin Earth liquid is simply not water would be inconsistent with the predictions of the first and fourth types of theories.

2.1 | Just superficial properties theories of natural kind concepts

A first possible type of theory holds that what is relevant to people's intuitions about natural kind category membership is just an entity's superficial properties. Since the liquid in the Twin Earth thought experiment has all of the relevant superficial properties, theories of this type predict that people's intuition should simply be *yes*, the Twin Earth liquid is water.

This first type of theory can be elaborated in several different ways. One possible view is that there is some specific set of superficial properties such that an entity will only be categorized as falling under the kind if it has *all* of those properties. A less stringent *just superficial properties* view might instead hold that an entity only needs a certain amount of the associated features to be categorized as a member of the kind. On the less stringent view, an entity might be categorized as a member of the kind, even if the entity is missing a few of the properties associated with the kind. For example, since green tea lacks the prototypical color associated with water, it would not be categorized as water according to the first view, but it might still be categorized as water according to the second.

Some specific *just superficial properties* views might also posit graded category membership. For example, consider a variation of the second "less stringent" view discussed previously

that also permits graded category membership. That view might predict that green tea is categorized as "sort of water," since it has some, but not all of the superficial properties associated with water.

The broader commonality among these more specific views is that only superficial properties determine the categorization of natural kinds. Although there are some important differences among various *just superficial properties* views, theories of this type all make the same prediction in Twin Earth cases. Because the Twin Earth thought experiment posits an entity with *all* of water's associated superficial properties, any *just superficial properties* view predicts that the intuition should be that the liquid is water.

2.2 | Just causal properties theories of natural kind concepts

A good deal of research suggests that the *just superficial properties* type of theory is, at best, incomplete. We now have overwhelming evidence that causal properties play a crucial role in natural kind categorization (e.g., Gelman, 2004; Keil, 1995). For this reason, that first type of theory has few, if any, contemporary defenders.

A second possible type of theory goes to the opposite extreme, suggesting that only causal properties determine natural kind categorization. This second type of theory holds that the role of representations of causal properties and superficial properties are importantly different, such that people only regard representations of causal properties as criteria of category membership.

One way of elaborating this type of theory would be to rely on the claim, introduced briefly above, that these two kinds of properties are represented in two very different formats. Perhaps the deeper causal properties are represented by a theory, the more superficial properties by a prototype. One could then claim that theories and prototypes have very different roles in natural kind categorization. It might be that people see the theory as the sole categorization criterion, but also use the prototype under time pressure or in situations of uncertainty. For instance, people might treat their theory of water as the sole categorization criterion, but people might also have a prototype of water that is helpful to get by, perhaps in the face of time pressure or lack of information, or one that plays a role in development (e.g., Keil, 1989).

By analogy, consider research on people's concept of prime numbers (Armstrong, Gleitman & Gleitman, 1983). People have a theory of prime numbers that provides a clear criterion for category membership, but people also appear to have a prototype of prime numbers (e.g., seven seems more "prototypically prime" than two). In this case, people's real criterion of category membership appears to be given by the theory, not the prototype (see also Gelman & Waxman, 2007).

Since this second type of theory says that when people evaluate the Twin Earth liquid, what is criterial is the deeper causal properties, it predicts that people's Twin Earth intuition should be simply *no*, the Twin Earth liquid is not water.

2.3 | Both together theories of natural kind concepts

A third type of theory rejects the hypothesis that only superficial or only causal properties feature in the categorization process. Instead, there might be a process of natural kind categorization that looks to both superficial properties and deep causal properties.

This third type of theory also encompasses a broad set of more specific views, so we can elaborate this third kind of theory in several different ways. One would be that there are two distinct

representations that each feed into one categorization process. For example, the categorization criterion for water might be given by the combination of a theory and prototype. That specific *both together* view would predict that the Twin Earth liquid is intuitively not water.

Another more radical *both together* hypothesis would be that there is one single representation of a given natural kind, and the superficial properties and causal properties are both relevant to that representation. The most extreme version of this view would hold that the categorization criterion is possession of all of the relevant superficial and causal properties. On that view, the Twin Earth liquid is not water since it lacks the relevant causal property. But a less extreme *both together* view might hold that the categorization criterion is possession of *some* of the relevant superficial and causal properties. Certain versions of that less extreme view might predict that the Twin Earth liquid would be judged as water.

The defining feature of this broader *both together* type of theory is that both superficial and causal properties feature in *one* categorization process. This third type of theory is actually compatible with the predictions of either of the first two types of theories. It might be that looking to both causal and superficial properties results in a judgment that the Twin Earth liquid is water, or that it is not water. Alternatively, insofar as the two criteria pull in opposite directions and some *both together* views allow graded category membership, those views might predict that people regard the Twin Earth example as a borderline case. Thus, some versions of *both together* views are also compatible with the judgment that the Twin Earth liquid is "sort of water."

2.4 | Each separately theories of natural kind concepts

A final type of theory posits two different processes of natural kind categorization. One process is guided by an entity's superficial properties, the other by an entity's deep causal properties. These two processes could then yield opposing conclusions about the very same case. Thus, a single entity might be categorized as a category member by one of these processes but as not a member by the other.

This type of theory is similar to the *both together* type in that it posits a role for both superficial and deeper causal properties in natural kind categorization. However, it differs from *both together* views in that it does not posit a single process that is influenced by both superficial and causal properties. Instead, the *each separately* type of theory posits *two* distinct categorization processes. One categorization process is guided by superficial properties; the other process is guided by causal properties. Thus, on this type of theory, a person could categorize the very same entity using two distinct processes and arrive at two different categorization intuitions.

One way of elaborating a more specific *each separately* view would be to hypothesize that people have both a theory and prototype of natural kinds. There would then be two distinct categorization processes: One that relies on the theory and one that relies on the prototype. This approach has been developed in important recent work by Machery (2009), Weiskopf (2009), Horst (2016), and others, and one might draw on this recent work to develop a more specific view along these lines. However, there might be other ways of elaborating an *each separately* view, such as positing that concepts like water have a natural and artifact kind categorization (e.g., Bloom, 2007).

This fourth type of theory opens up a new possibility: The Twin Earth liquid could be judged to be water in one sense but also to be not water in another sense.

3 | TWIN EARTH INTUITIONS AND THEORIES OF SEMANTICS

Judgments about the Twin Earth thought experiment may also serve as evidence bearing on various theories of the semantics of natural kind terms. The relation between people's ordinary Twin Earth intuitions and these semantic theories is more complex than the relation between the intuitions and theories in cognitive science. Although there is a strong consensus that empirical facts about people's intuitions are directly relevant to cognitive-scientific theories of natural kind categorization, their relevance to questions in semantics is more controversial. Some philosophers argue that empirical facts about people's intuitions are relevant to semantic questions (e.g., Corcoran, 2018), while others argue that they are not (e.g., Deutsch, 2015). Our paper cannot possibly settle this broad meta-philosophical dispute. Our more modest aim is to contribute to debates about what people's intuitions about natural kind categorization actually are.

We divide the conceptual space in a similar way to the division we used above for cognitive theories. We consider four broad types of theories about the semantics of natural kinds: The reference of a natural kind term might be determined by (a) just superficial properties; (b) just causal properties; (c) both kinds of properties together; or (d) each kind of property, separately.

3.1 | Just superficial properties theories of natural kind semantics

A first type of theory posits that the extension of a natural kind term should be understood entirely in terms of superficial properties. For example, the word "water" might pick out entities that share certain superficial properties, regardless of whether they also share any deeper causal properties. On this first type of theory, the correct answer is that the Twin Earth liquid is water.

On a plausible reading of Putnam (1975) and Kripke (1972/1980), this type of theory was the target of the Twin Earth thought experiment. The purported fact that the Twin Earth liquid seems not to be water is taken as evidence against theories of this type.

3.2 | Just causal properties theories of natural kind semantics

On a second type of theory, the reference of a natural kind term is determined only by deeper causal properties. The term "water" then picks out entities that share certain deeper causal properties, regardless of whether those entities also share any superficial properties. A common interpretation of Putnam's (1975) position falls under this type of view (for further discussion, see Pessin & Goldberg, 1996).

Theories of this second type could be elaborated within several different meta-semantic frameworks. Defenders of causal-historical frameworks would say that the reference of "water" is H_2O in virtue of the term's causal history (Kripke, 1972/1980; Putnam, 1973, 1975), while defenders of descriptivist frameworks would say that the reference of "water" is H_2O in virtue of the term's descriptive content (e.g., Jackson, 2003, p. 60; Chalmers, 2002). Specific variants of these latter theories might have superficial properties serve a reference-fixing function (e.g., Jackson, 2003).

Although there are important differences among these meta-semantic frameworks, the important unifying feature of *just causal properties* views is that is that, ultimately, a term like "water" picks out a set of entities that share certain deeper causal properties, but that may not also share any superficial properties. On such views, the correct answer is that the Twin Earth liquid is not water.

3.3 | Both together theories of natural kind semantics

A third type of theory is that the referent of a natural kind term is given by the combination of its deep properties and superficial ones. As with the previous types of theories, this *both together* theory may be elaborated in multiple ways. One would be to say that "water" refers to anything that has *most* of the relevant properties, where the relevant properties include both deep causal ones and superficial ones.

On a given *both together* view, the correct response to the Twin Earth case depends on the weight assigned to causal and superficial properties. A variant that places more weight on causal properties might hold that the Twin Earth liquid is water, while a variant that places more weight on superficial properties might hold that the Twin Earth liquid is not water. One that places moderate weight on both might hold that the Twin Earth liquid is a borderline case, best understood as "sort of water."

Thus, on *both together* views, the correct answer is one of the following: The Twin Earth liquid is water; it is not water; or there is some vagueness such that it is "sort of water."

3.4 | Each separately theories of natural kind semantics

A final type of theory is that each natural kind term has two different senses, each of which picks out a different set of entities. One sense picks out entities that share certain superficial properties (e.g., clear, potable, tasteless); the other would pick out entities that share a deeper causal property (e.g., being H_2O).

This *each separately* type of theory can be clarified by contrast to the previous one. On any *both together* view, each natural kind term has a single sense, and both superficial and causal properties are relevant to that one sense. By contrast, on any *each separately* view, natural kind terms have two different senses. One sense involves just superficial properties; the other sense involves just causal ones.

In almost all cases, these two senses would yield the same verdict. Liquids in our world that are composed of H_2O are typically clear, potable, and so forth—and vice versa. These cases do not allow one to distinguish the *each separately* type of theory from the others. When a liquid has both sets of properties, an *each separately* view says that there is no sense at all in which it is not water. The other three types of theories return the same verdict.

The distinctive predictions of each separately views only come out in the more unusual case in which an entity has the superficial properties associated with a natural kind, but not the causal ones. In that kind of case, an each separately view says something that other types of theories do not: There is one sense in which the entity is water, but another sense in which it is not. Twin Earth cases are therefore an ideal test for this type of theory.

In short, on this final type of theory, the correct answer is that there is a sense in which the Twin Earth liquid is really water, and also a sense in which the liquid is not really water at all.

4 | EXISTING WORK ON THE ROLE OF SUPERFICIAL PROPERTIES

The standard Twin Earth intuition suggests that superficial properties play no role in categorization (e.g., the Twin Earth liquid is superficially *exactly like* water, yet it is not water). Nevertheless, a number of studies suggest that superficial properties do play some role in natural kind categorization. Section 4.1 presents some of those prior studies.

A question arises about how to best understand these prior results. Section 4.2 suggests that recent work in an apparently unrelated area of inquiry might shed further light on how exactly superficial properties figure in natural kind categorization. Drawing on work on "dual character concepts," we hypothesize that Twin Earth intuitions represent a dual character pattern of judgment.

4.1 | The relevance of superficial properties

Within existing work, there have been a number of findings suggesting that causally central properties are not people's sole criteria for natural kind membership (see, e.g., Braisby, Franks & Hampton, 1996; Genone & Lombrozo, 2012; Jylkka, Railo & Haukioja, 2009; Nichols et al., 2016). For example, Malt (1994) demonstrates that the presence of H_2O in a liquid is not the only property that plays a role in categorization as water. Although tea is presumably not water, participants judged it as 91% H_2O. By contrast, although salt water is presumably water, participants judged it as 83% H_2O. Drawing on this finding, Malt argues that that something besides deeper causal properties must play a role in categorization. Various replies have defended the exclusive significance of causal properties against these results (e.g., Abbott, 1997; Ahn et al., 2001).

Still, it seems that a central piece of evidence supporting the significance of causal properties, which we call the "just causal properties" view, is the type of judgment exemplified and popularized by Twin Earth cases. The liquid in the thought experiment is identical to H_2O with respect to *every* superficial, observable property. Nevertheless, the liquid's distinct underlying causal structure distinguishes it from water. Thus, if people do in fact have the intuition that this liquid is not water, this intuition would provide strong reason to adopt the *just causal properties* view.

However, recent experimental work suggests—perhaps surprisingly—that people do not actually have the standard philosophical intuition about Twin Earth cases. In one recent study, Corcoran (2018) presented participants with three Twin Earth cases (involving water, gold, or tigers). In each case, the entity was described as having all of the relevant superficial properties, but lacking the relevant deeper causal properties. Participants were asked to agree or disagree with a categorization statement about the Twin Earth entity (e.g., "The liquid is water"). Mean responses fell towards the middle of the rating scale, with participants showing a strikingly bimodal response pattern. This suggests that at least some people categorize entities using more than solely deep causal properties.

4.2 | Dual character

To gain some insight into these results, we explore what might seem like an unrelated phenomenon: Dual character concepts (see Knobe, Prasada & Newman, 2013). Each dual character concept appears to be associated with two distinct sets of criteria for category membership. One involves the superficial properties, while the other involves abstract values.

As an example, consider the concept ARTIST. Now imagine a person who creates paintings for a living but who has no real interest in creating work of deep aesthetic value and is simply trying to make money. When evaluating such a person, experimental participants agree that both:

1. There is a sense in which this person is an artist.
2. Ultimately when you think about what it really means to be an artist, you would have to say that this person is not truly an artist.

Thus, the very same entity can be seen as falling under the concept in one sense but not in another.

Some existing work suggests that this dual character pattern might extend to natural kind concepts.[1] In other words, it might be that people are inclined to agree that an entity falls under a natural kind in one sense, but not in another sense. For example, Machery and Seppälä (2011) find that some experimental participants are inclined to agree both with the statement "In a sense, tomatoes are vegetables" and with the statement "In a sense, tomatoes are not really vegetables."

We predict that people have a "dual character" intuition about Twin Earth scenarios. For instance, in cases in which a liquid shares all superficial properties of water, but not the underlying causal property, we predict that people will agree with both of the following statements:

1. There is a sense in which the liquid from Twin Earth is water.
2. Ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid from Twin Earth is not truly water at all.

On this prediction, people explicitly acknowledge two senses of the term, concluding that the Twin Earth entity falls under the concept in one of the senses but not in the other.

Of course, in many cases people do not have the opportunity to use this more complex statement. In cases in which people are confronted with a binary choice about category membership, they have to use one or the other of the two criteria. In such a case, we predict that context will affect which criteria they use. For instance, imagine trying to determine whether a liquid counts as "water" in two different contexts. One is a scientific context in which chemists are seeking experimental material ("Bring me a 12 mL sample of water"); the other is a social context in which a mother is giving a warning to her child ("Don't play in the water"). These different contexts suggest the relevance of different categorization criteria. We predict that in a binary choice, people would be more inclined to categorize the liquid with respect to the deeper causal properties in the scientific context and with respect to the superficial properties in the social context.

The dual character pattern we predict here is most naturally explained by the *each separately* family of theories. However, the pattern of people's judgments in Twin Earth cases is just one of the many sources of evidence that bear on these broader theories, and one might well think that evidence from some other source gives us reason to prefer one of the other theories. We will return in Section 6 to the question as to whether any other type of theory might be able to explain the dual character pattern.

5 | EXPERIMENTS

The present studies aim to determine whether people's natural kind categorization judgments display a dual-character pattern. Two of the studies reported here used Twin Earth cases. In these studies, participants received the classic case involving water (Putnam, 1975) and closely parallel cases involving tigers and gold (Kripke, 1972/1980). To eliminate researcher degrees of freedom, we did not write the vignettes describing these cases ourselves. Instead, we used the exact wording of the materials from a previous paper designed for a different experimental purpose, without knowledge of our present hypothesis (Corcoran, 2018). In other words, the present studies use the exact cases first

[1] Early work on the dual character pattern of judgment focused on value-based concepts (like the concept ARTIST), and a question therefore arises as to how to understand the relationship between these value-based concepts and natural kind concepts. This question is outside the scope of the present paper, but see Newman and Knobe (2018) for an argument, drawing on the results reported here, that value-based concepts are actually far more similar to natural kind concepts than they might at first appear to be.

introduced to support the view opposed to what we predict people will think and also use a way of writing out those cases introduced by another researcher who was not aware of our hypothesis.

Experiments 1a and 1b test whether participants endorse the standard philosophical intuition about Twin Earth cases, or whether they instead assent to the claim that there is one sense in which the entity is a member of the natural kind category but also another sense in which it is not. Experiment 2 tests whether natural kind categorization is affected by situational context. Experiments 3a and 3b extend this test by replacing philosophical Twin Earth cases with a real-world example (about genetically modified organisms) and also by looking at a more highly educated population (graduate students at elite universities).

5.1 | Experiment 1

Participants were presented with different versions of "Twin Earth" scenarios. Specifically, all participants read about entities that lacked the deeper causal properties associated with a particular natural kind (e.g., a liquid with a different chemical structure than water). In one condition the entity had all of the superficial properties associated with the kind, while in the other condition the entity lacked those properties.

If the standard philosophical intuition is shared, there should be no difference in ratings between conditions; in both cases, participants should say that the entity is not a member of the natural kind. By contrast, if the dual-character prediction is correct, participants should show a more complex pattern of judgments. When the entity lacks the superficial properties, participants should be inclined to say that the entity is not a member of the natural kind in any sense. When the entity has the superficial properties, participants should be inclined to say that the entity is a member of the kind in one sense, but is not a member in another sense.

Experiment 1a tests this prediction by presenting participants with a forced choice between statements. In Experiment 1b, participants were given Ą two statements—one affirmative (is a member of the category) and one negative (is *not* a member of the category)—and had an opportunity to separately express agreement or disagreement with each.

5.1.1 | Experiment 1a

Method

Participants Six-hundred participants were recruited from Amazon's Mechanical Turk (56% male, 43% female, 0% non-binary, mean age = 34).

Materials and Procedure Each participant was presented with one vignette presented in a 3 (Kind: Gold, tigers, or water) × 2 (Vignette Structure: Same appearance, different appearance) between-subjects design. In the same appearance conditions, participants read vignettes from Corcoran (2018). Those vignettes described an entity (e.g., a liquid) that had all of the same superficial properties as an entity on Earth (e.g., drinkable, clear), but a different causal property (e.g., not H_2O). In the different appearance condition, participants read vignettes in which the entity did not have any of the superficial properties as an entity on Earth (e.g., is not potable and does not look like water), and also had a different causal property (see the Appendix at https://osf.io/8953f/ for vignettes, all of which are taken unchanged from Corcoran, 2018).

Participants then received three statements and were asked to indicate which one they agreed with most. For example, the question for water was: With which of the following do you most agree?

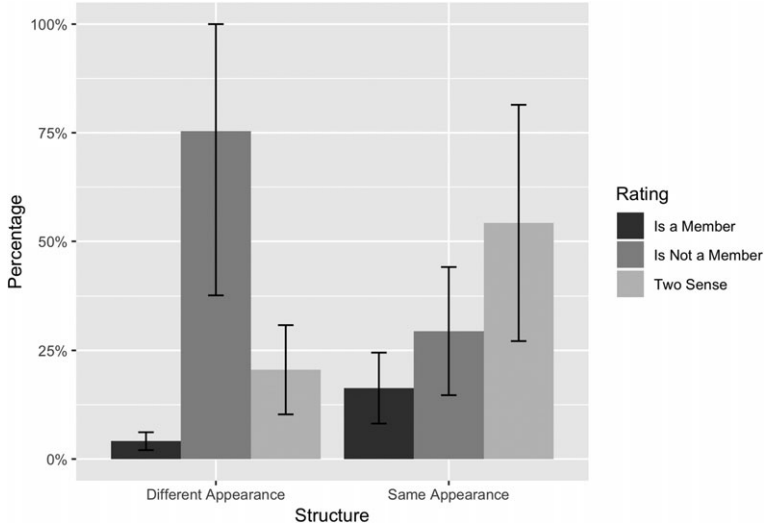

FIGURE 1 Percentages of participants choosing each statement (Same Appearance vs. Different Appearance), collapsing across Kind (gold, tiger, water). Error bars indicate 95% confidence intervals

1. The liquid from Twin Earth is water.
2. The liquid from Twin Earth is not water.
3. There is a sense in which the liquid from Twin Earth is water, but ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid from Twin Earth is not truly water at all.

These three options were presented in a random order. After responding to the forced choice question, participants responded to two comprehension check questions. All vignettes and questions are listed in full in the Appendix (https://osf.io/8953f/).

Results

Three hundred and ninety-one participants correctly responded to the two comprehension check questions. Analyses were conducted on these participants. The percentage of participants choosing each option is shown in Figure 1. (Data from all experiments are available on the Open Science Framework: https://osf.io/f44hq/.)

We analyzed the data using two hierarchical binary logistic regression models. In both models, the dependent variable dichotomized participants' responses as either "Is *not* a member" or another response (either "Is a member" or "Two Sense"). In the first model we entered as predictors Vignette Structure (different appearance vs. same appearance) and two dummy codes for the Kind (gold, water). In the second model we also included two interaction terms (gold × vignette structure, water × vignette structure). The comparison of these models indicated that Vignette Structure did not significantly interact with Kind, $X^2(2, N = 391) = 2.63$, $p = 0.269$.

The results showed a main effect of Vignette Structure ($B = -1.97$, $SE = 0.24$, $p < 0.001$, odds ratio [OR] = 7.18, where participants were less likely to choose the "Is *not* a member" statement in the Same Appearance condition than in the Different Appearance condition.[2] Moreover, this pattern significantly replicated for all three Kinds (gold, X^2 (1, $N = 113$) = 19.34, $p < 0.001$; tiger $X^2(1, N = 144) = 20.10$, $p < 0.001$); water, $X^2(1, N = 134) = 38.10$, $p < 0.001$. Finally, as seen in

[2] We also conducted an analysis on all participants, including those who failed check questions. An inclusive analysis (excluding no participants) also reveals no difference between the two models, $X^2(2, N = 601) = 4.61$, $p = 0.100$. There was also a main effect of Vignette Structure ($B = -1.05$, $SE = 0.293$, OR = 0.35, $p < 0.001$).

TABLE 1 Percentages of participants (correctly responding to both comprehension questions) choosing each statement (Same Appearance, Different Appearance)

	Two sense	Is a member	Is *not* a member
Same Appearance—Different causal structure			
Gold	0.68***	0.06***	0.26
Tiger	0.42*	0.21*	0.37
Water	0.54**	0.21*	0.25
Different Appearance—Different causal structure			
Gold	0.19*	0.11***	0.69***
Tiger	0.25	0***	0.75***
Water	0.15**	0.04***	0.81***

Notes. Asterisks indicate significance via a binomial comparison to chance (0.33).
*$p < 0.05$.; **$p < 0.01$.; ***$p < 0.001$.

Table 1, the Two Sense statement was the most popular response in the Same Appearance condition, while the non-member statement was the most popular in the Different Appearance condition.

5.1.2 | Experiment 1b

Method

Participants One hundred and eighty-two participants were recruited from Amazon's Mechanical Turk (62% male, 36% female, 2% non-binary, mean age = 34).

Materials and procedure The design of Experiment 1b was identical to that of Experiment 1a, except that participants responded to two scaled rating questions rather than one forced choice question. For example, for water, participants were asked to rate their level of agreement with two statements:

1. There is a sense in which the liquid from Twin Earth is water.
2. Ultimately, if you think about what it really means to be water, you would have to say there is a sense in which the liquid from Twin Earth is not truly water at all.

Participants rated both statements on a scale from 1 (disagree) to 7 (agree).

Results

The mean response for each question across vignettes is displayed in Figure 2.

Results for the member statement and non-member statement were analyzed separately. For each statement we conducted a 2 (Vignette Structure: Different appearance, same appearance) × 3 (Kind: Gold, tiger, water) analysis of variance (ANOVA).

For the member statement, there was a main effect of vignette structure, such that participants in the same appearance conditions agreed more strongly ($M = 4.73$, $SD = 1.53$) than participants in the different appearance conditions ($M = 2.92$, $SD = 1.86$), $F(1, 176) = 51.90$, $p < 0.001$, $\eta_p^2 = 0.23$. There was no effect of Kind and no interaction.

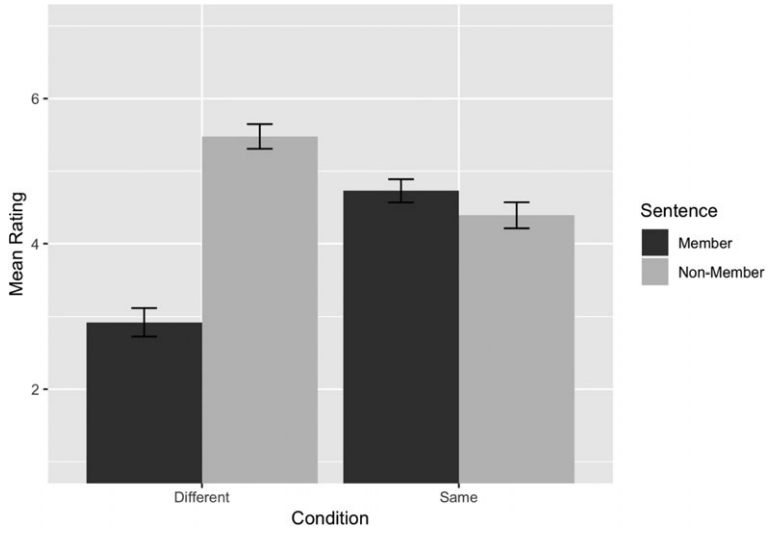

FIGURE 2 Mean ratings for each statement, collapsing across Kind. Error bars indicate *SE*

For the non-member statement, there was a main effect of vignette structure, such that same appearance condition participants agreed less strongly ($M = 4.39$, $SD = 1.72$) than participants in the different appearance condition ($M = 5.48$, $SD = 1.6$), $F(1, 180) = 19.04$, $p < 0.001$, $\eta_p^2 = 0.10$. There was no effect of Kind and no interaction.

To further explore participants' responses in the same appearance condition, we then conducted one-sample *t* tests comparing responses on each of the questions to the scale midpoint (4). Results indicated that ratings we significantly higher than the midpoint both on the member statement, $t(91) = 4.553$, $p < 0.001$, and on the non-member statement, $t(91) = 2.179$, $p = 0.032$. In other words, participants agreed both with the statement that there is a sense in which the entity is a member of the kind *and* with the statement that there is a sense in which the entity is not a member of the kind.

Figure 3 presents the paired ratings of each Same Appearance participant for the Member and Non-Member statements. Participants who endorsed the standard intuition are those in the upper-left quadrant. As Figure 3 indicates, most participants rejected the standard intuition. Moreover, a substantial number of participants rated *both* statements above the scale midpoint.

Discussion

Recall the standard philosophical intuition: The Twin Earth liquid is not water. Two studies showed that people's ordinary responses do not conform to this standard intuition. Most participants did not select the response consistent with the standard intuition in a forced choice, and most participants agree that there is a sense in which the (Same Appearance) Twin Earth liquid is water.

Although there is some variation among the participants' judgments, we interpret the overall results as best supporting a dual-character pattern of judgment: When an entity lacked both the underlying causal properties and superficial properties, participants were inclined to say it was not a member of the kind in any sense, but when an entity lacked the underlying causal properties but shared the superficial properties, participants were largely inclined to say it was not a member of the kind in one sense, but was a member in another sense. This dual-character response was the most popular response in a forced-choice task (see Table 1), participants' mean ratings suggest overall agreement with both senses of category membership (see Figure 2), and most individual participants rated agreement with each sense at the scale midpoint (4) or above (see Figure 3).

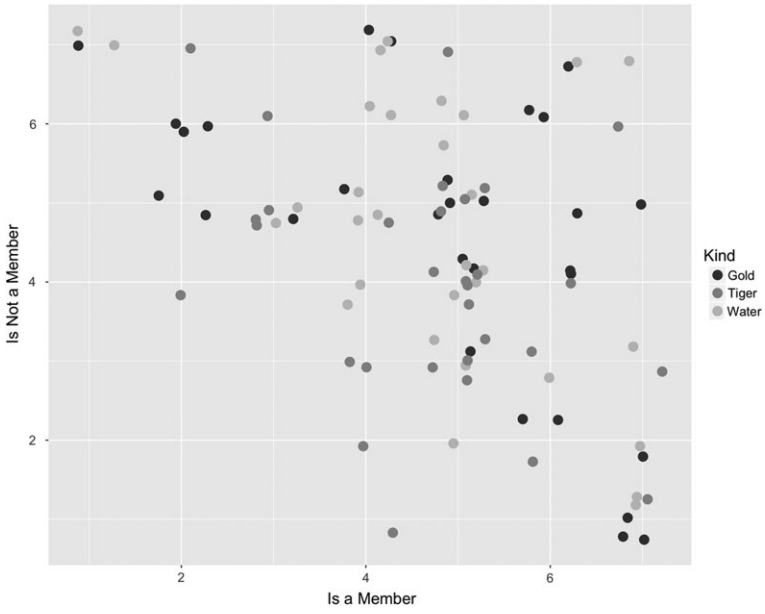

FIGURE 3 Scattterplot (with jitter) of responses to Member and Non-Member statements in experiment 1b

The most straightforward interpretation of this dual-character pattern of judgment seems to support *each separately* theories of people's ordinary natural kind concepts, according to which natural kind concepts are seen in terms of two sets of criteria, one involving deep, causal properties and another involving superficial properties. Similarly, the most straightforward interpretation of the results seems to support *each separately* theories of semantics, according to which natural kind terms have two different senses. However, there might be ways to interpret the results such that they are consistent with other theories. We explore these possibilities in Section 6.

We think that the dual-character interpretation is the best interpretation of the overall pattern of results, but there is still something puzzling about the disagreement among participants. The majority of participants choose the dual-character statement, but there are others who judge that the Twin Earth liquid is simply water or simply not water.

There are different plausible hypotheses about this disagreement. One possibility that is consistent with the dual-character interpretation is that part of the disagreement arises from the effect of *context.* Perhaps, for dually characterized concepts, context can suggest which set of categorization criteria is more relevant. For example, in a more scientific context people might be more inclined to judge that the Twin Earth liquid is water (e.g., if we are determining whether the liquid is a suitable sample of water in a science laboratory). But participants might be more inclined to judge that the liquid is *not* water in a social or legal context (e.g., in the context of a local ordinance that prohibits filling unapproved pools with water). In such context-rich examples, the context can suggest one sense of categorization criteria that is more relevant, leading people to make judgments that reflect only that set of criteria.

The standard Twin Earth thought experiment tested in Experiments 1a and 1b has much less context. But in those cases, some participants might have nevertheless *guessed* which context they were in and which set of categorization criteria was more relevant to that context. If so, we would expect that some participants might judge that the Twin Earth liquid simply is water and others might judge that it is simply not water. We test this possibility in the next experiment.

5.2 | Experiment 2

The previous experiments allowed participants to express two seemingly opposed views, revealing a dual-character pattern of judgment in Twin Earth cases. Specifically, the results thus far indicate that when an entity shares the superficial properties associated with a natural kind but lacks the associated casual properties, people are inclined to categorize the entity as a category member in one sense, but not a category member in another sense.

Although most participants endorse the dual-character statement, Experiment 1b revealed some disagreement about the Twin Earth case. Our hypothesis about this disagreement was that category membership judgments can be affected by contextual information that makes one or another set of properties seem more relevant. For instance, would participants be less inclined to categorize the Twin Earth liquid as water in the purely scientific context of a chemistry class? What about in a more practical context in which a town has a rule prohibiting residents from creating unapproved pools of water? We predict that which set of properties is most relevant (deep causal or superficial) can vary with context and that when participants are forced to choose whether the entity is or is not a natural kind category member, they will categorize the entity in line with the set of properties that the context indicates as most relevant.

5.2.1 | Method

Participants

Four hundred and fifty-six participants were recruited from Amazon's Mechanical Turk (62% male, 38% female, 0% non-binary, mean age = 33).

Materials and procedure

Participants received one of the Twin Earth vignettes (gold, tiger, water) and then were given information about one of three possible contexts: Scientific, legal, or neutral.

Participants in the scientific context conditions were told that a science department in a university had a rule stating that all students must be provided with certain objects (gold, tigers, or water) for their science practical testing. Participants in the legal context conditions were told that a town had a rule stating that certain objects (gold, tigers, or water) cannot be used for certain purposes without approval (housing additions, pet adoption, home pool creation). Participants in the neutral context were given no information about the context (see the Appendix at https://osf.io/8953f/ for full materials).

In all conditions, participants were then told there is a controversy about the entity's category members. Participants rated their agreement with a statement about category membership. For example, in the water conditions, the statement was "The liquid from Twin Earth is water" where 1 (disagree) and 7 (agree). Full vignettes and questions are listed in Appendix S1.

5.2.2 | Results

Mean ratings by condition are displayed in Figure 4. The data were analyzed using a 3 (Context: Science, neutral, legal) × 3 (Kind: Gold, tiger, water) ANOVA. There was a main effect of context, $F(2, 455) = 9.94$, $p < 0.001$, $\eta_p^2 = 0.043$. There was no effect of kind and no interaction (both Fs < 1). Post-hoc Tukey's tests showed that participants were more inclined to rate the object as a member of the category in the legal than in the science context, $p < 0.001$. Participants were also more

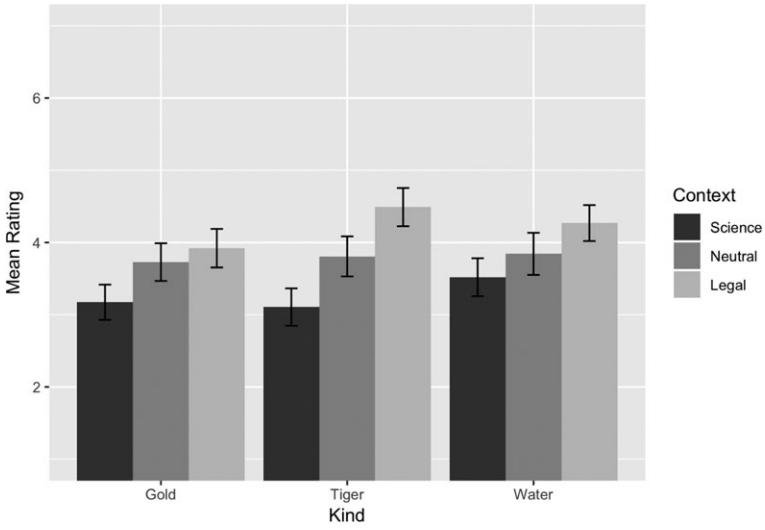

FIGURE 4 Mean ratings of kind by context. Higher ratings indicate categorization of the entity as a member of the natural kind. Error bars indicate standard error

inclined to rate the object as a member of the category in the neutral than in the science context, $p = 0.045$. The neutral and legal context ratings were not significantly different, $p = 0.109$.

5.2.3 | Discussion

Participants' judgments about category membership depended on context. In the science context, participants focused more on the deeper causal properties, while in the legal context, participants focused more on the superficial properties. The neutral context was intermediate between the two.

This result provides further evidence that people have distinct sets of criteria that determine category membership. Which set of criteria is employed depends on the particular context.

5.3 | Experiment 3

Studies 1 and 2 employ Twin-Earth style thought experiments. Since these have been taken as paradigmatic examples in support of the standard philosophical intuition, these studies provide evidence against the claim that people hold that intuition even in the cases introduced to support it. Instead, they provide evidence for the claim that people's judgments show a dual character pattern.

Although the classic Twin Earth thought experiments have been seminal examples, it might be thought that they are overly philosophical or esoteric. For this reason, Experiment 3 uses more realistic cases.

Finally, one might worry that the dual-character intuition arises only because participants fail to think clearly and carefully about the questions. For this reason, this final experiment was conducted on two different populations. Experiment 3a uses an online sample, while Experiment 3b turns to a sample of graduate students from elite universities.

5.3.1 | Experiment 3a

Methods

Participants One hundred and fifty participants were recruited from Amazon's Mechanical Turk (57% male, 43% female, 0% non-binary, mean age = 33).

Materials and procedure All participants read a vignette about genetically modified salmon, fish whose genes have been altered to enable them to grow at faster rates:

> The Maxwell Laboratory has made great progress researching fish genetics. They have discovered how to modify the genes of salmon in order to enable the fish to grow year-round instead of only during the summer months. These genes enhance speed of growth but they do not affect any other qualities. The laboratory's fish are identical in all other observable properties to salmon. These properties include appearance, size, taste, and other markers that distinguish salmon from other similar fish.
> If one were to perform a genetic analysis of one of the laboratory's fish, however, one would find the fish does not contain the genes of salmon. Instead, the laboratory fish contains the modified genes.
> The modified fish and salmon are completely indistinguishable and interchangeable outside of the laboratory. The laboratory issues a report stating that while the fish do not belong to the same scientific category as familiar salmon, this difference is immaterial for any purpose other than scientific classification.

Participants were randomly assigned to receive information about one of three contexts (as in Experiment 3). In the science context, participants received a story about fish used for testing in a science laboratory. In the legal context, participants received a story about fish sold at a farmer's market. Participants in the neutral context were given no information about the context (see Appendix S1).

All participants rated their agreement with the category membership statement "The fish from the laboratory are salmon" on a scale from 1 (disagree) to 7 (agree).

Results

Mean ratings by context are displayed in Figure 4. A one-way ANOVA found a significant effect of context, $F(2, 149) = 3.36$, $p = 0.028$, $\eta_p^2 = 0.047$. Post-hoc Tukey's tests showed that participants were more inclined to rate the fish as salmon in the legal context than in the science context, $p = 0.021$. Neutral context ratings did not differ significantly from the legal context ratings, $p = 0.259$, or the science context ratings, $p = 0.505$.

5.3.2 | Experiment 3b

Methods

Participants One hundred and ninety-three participants were recruited from elite graduate programs (50% male, 47% female, 3% non-binary, mean age = 27). To recruit participants, we emailed department administrators from a diverse selection of graduate programs (Anthropology, Economics, Geology/Geophysics, Neuroscience/Neurobiology, Political Science/Government, Sociology, and Statistics) at elite universities (Harvard, Princeton, Stanford, and Yale University). We planned to continue emailing new departments until any round of emails brought our total participant number past 150. Our first round of emails, to seven departments at four universities, recruited 193 participants. See Table 2 for participants' graduate degree universities and departments.

Materials and procedure The materials and procedure are identical to that described in Experiment 3a.

TABLE 2 Number of participants by graduate degree university and department

	Harvard	Princeton	Stanford	Yale	Other	Total
Anthropology				1		1
Economics		16	18	15		49
Geology/Geophysics				5		5
Neuroscience/Neurobiology	22			16	1	39
Political Science/Government	29	17		1		47
Sociology	21		12			33
Statistics				3		3
Other	3	1	1	3	8	16
Total	75	34	31	44	9	193

Notes. "Other" includes no response and responses that were ambiguous between categories (e.g., "Public Policy").

Results

Mean ratings by context are displayed in Figure 5. A one-way ANOVA found a significant effect of context, $F(2, 190) = 6.60$, $p = 0.002$, $\eta_p^2 = 0.065$. Post-hoc Tukey's tests showed that participants were more inclined to rate the fish as salmon in the legal context than in the science context, $p = 0.001$. Neutral context ratings did not differ significantly from the legal context ratings, $p = 0.065$, or the science context ratings $p = 0.308$.

Finally, we considered the online and graduate student sample together, conducting a 2 (Population: Online, graduate) × 3 (Context: Science, neutral, legal) ANOVA. There was a main effect of context, $F(2, 337) = 9.86$, $p < 0.001$, $\eta_p^2 = 0.055$. There was no main effect of population, $F < 1$, and no interaction, $F < 1$.

We conducted two analyses to compare the attention and effort of the graduate student and online (MTurk) participants. First, we compared the percentage of participants correctly answering a reading comprehension check question. All graduate participants answered the check question correctly, and all but two of the online participants answered the question correctly, Fisher's exact test $p = 0.19$. As

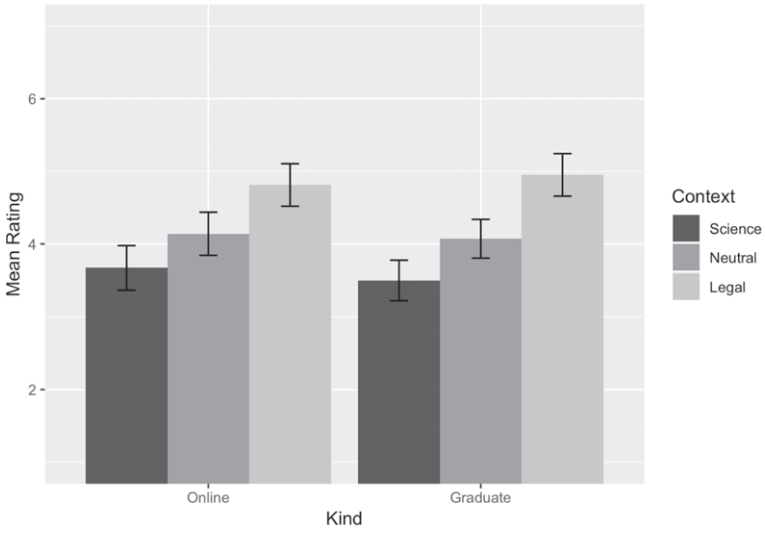

FIGURE 5 Mean ratings by context. Higher ratings indicate categorization of the entity as a member of the kind. Error bars indicate standard error

a measure of effort we compared average survey time. Excluding outliers who took over 20 min to complete the study, the online (MTurk) participants spent on average 2.76 min ($SD = 1.82$), while graduate students spent on average 3.40 min ($SD = 2.01$), t (1, 324) = 2.98, $p = 0.003$. Including those who took over 20 min only increases the graduate student average, and all online (MTurk) participants spent less than 20 min. Thus, while graduate students and online participants are both paying sufficient attention to correctly answer comprehension questions, graduate students appear to be attending to the question with increased consideration.

Discussion

Experiment 3 examined participants' intuitions in a more realistic scenario. Once again, there was an effect of context on participants' judgments of category membership. In the scientific context participants were less inclined to categorize the entity as a member of the natural kind. In the legal context they were more inclined to categorize it as a member. The neutral condition was intermediate.

The results also suggest that the dual-character context effect is not simply a result of careless or inattentive thinking. The same effect arose in both the online and graduate student populations. Thus, even participants with extremely high levels of education categorize these entities in a way that is sensitive to context, in line with the dual-character prediction.

6 | IMPLICATIONS

Our experiments reveal that people's Twin Earth judgments display a dual-character pattern. People are inclined to endorse the dual-character statement over a statement expressing the classic Twin Earth intuition; they endorse both conjuncts of the dual-character statement; and when they are forced to make a binary choice, context guides which option they select. Moreover, this context effect arises in Twin Earth cases alongside more realistic cases and across ordinary participants and "elite university" graduate students.

In this section we address how these results bear on deeper questions, exploring both cognitive and semantic implications.

6.1 | Cognitive implications

With respect to the different cognitive hypotheses about natural kind categorization processes, one natural interpretation is that our results provide evidence for *each separately* views. Those views posit two distinct natural kind categorization processes, such that one relies on superficial properties and the other on deeper causal properties. These two processes could easily account for people's two, seemingly opposed intuitions in Twin Earth cases.

There are a number of ways to further elaborate such an *each separately* view, but one obvious approach would be to draw on existing frameworks according to which people can associate a single concept with a number of different criteria in different representational formats (e.g., Machery, 2009; Weiskopf, 2009). Perhaps a natural kind's set of deep causal properties is represented by a theory and the set of superficial properties by a prototype. Each of those is associated with a distinct categorization process. When a context invites categorization by both of those processes, people can report seemingly contradicting verdicts (e.g., in Twin Earth cases). When a context invites only one categorization, people often look to context to indicate which representation is most relevant.

Although it is most straightforward to understand the results as supporting *each separately* views, we can also consider how they might be consistent with the *just causal properties* views. For

example, perhaps the participants simply did not interpret the stimuli in the way we intended. In Experiment 1a, participants endorse this statement about the Twin Earth liquid (emphasis added):

> There's a sense in which the liquid from Twin Earth is water, but ultimately, if you think about what it really means to be water, you'd have to say there's a sense in which the liquid from Twin Earth is not truly water at all.

One possible interpretation of this sentence would be that the liquid can be called "water" in some loose non-literal sense but does not count as "water" in any strictly literal sense.

This particular interpretation seems less plausible when we consider other dual-character patterns of judgment. Consider a mother who has no real feelings for her children and only takes care of them because she is concerned with being seen to be caring. In cases of this type, studies show that people endorse a statement of the form: "There is a sense in which she is clearly a mother, but ultimately, if you think about what it really means to be a mother, you would have to say that she is not a mother at all" (Knobe et al., 2013). In such an endorsement, participants do not mean to say there is no literal sense in which she is a mother. Rather, they mean that there is a literal sense in which she is a mother, but also another, deeper sense in which she is ("ultimately") not a mother.

We can also consider how the results might be consistent with certain *both together* views. Perhaps in Experiment 1a, participants chose the dual-character statement because it allows participants to express an intermediate degree of category membership. For a potentially analogous case, consider a man who is 5′11″. Rather than saying that he is tall or that he is not tall, perhaps it is more expressive to say that there is a sense in which he is tall and a sense in which he is not tall.

To test this interpretation of Experiment 1a, future experimental work could test whether people endorse dual-character statements in a case that clearly seems to involve one categorization process. For example, consider again the question as to whether a man who is 5′11″ is "tall." In a case like this, it seems that there are not two categorization processes. Instead, it seems like there is only one categorization process such that this is a borderline case. The key question is whether people in this case would endorse a sentence like: "There is a sense in which he is really tall, but ultimately there is a sense in which he is really not tall." If people do endorse a sentence like this, our Experiment 1a would provide less evidence against some *both together* views.

It is less obvious how to square this *both together* interpretation with the results of Experiment 1b. In that experiment, participants agreed with both conjuncts of the dual-character statement (Is a member and Is Not a member) at levels just above the scale midpoints. This result seems to conflict with the *both together* thesis that there is just one sense of category membership. Of course, the strongest evidence against *both together* views would be *strong* agreement with both statements (Is and Is Not a member) at very high scale levels. We found moderate agreement with both statements, providing relatively weaker evidence against this *both together* interpretation.

Finally, consider how this *both together* interpretation might be consistent with the context results. The *both together* view posits that both causal properties and superficial properties are relevant to *one* categorization process. For such a view to account for the context effect, it might add that the relative weight of these properties varies based on context. Perhaps concepts like HEALTHY have this feature (e.g., Sassoon, 2013). In general, both low sugar content and high vitamin content are relevant properties. Across different contexts, the relative weights of these properties might differ. Thus, in one context orange juice might be judged as healthy, but in another as not healthy. A *both together* view might posit that something similar is true of natural kind concepts.

Ultimately, we do not take our results to *settle* the question of which of these types of theories is correct. Instead, the core finding of the present studies is simply that people's intuitions about natural kind concepts show a dual character pattern. Although we take this result to provide some evidence for *each separately* views, further theorizing and research might reveal this dual character pattern to be compatible with versions of some other types of theories. Regardless of which type of theory turns out to be correct, our claim is that any view should be elaborated to account for the dual character pattern of judgment, rather than for the standard philosophical intuition.

6.2 | Semantic implications

With respect to the different hypotheses about semantics, we understand the story to be more complicated. While there is a strong consensus that empirical data are straightforwardly relevant to the question discussed in the previous section, a great deal of controversy remains about the role of empirical data in semantic theory. Nevertheless, it is plausible that empirical facts about people's ordinary Twin Earth intuitions have at least some relevance for research on semantic questions. The response that the Twin Earth liquid is not water has traditionally been described as involving an "intuition." Some describe it as the "standard intuition" (Laurence, Margolis & Dawson, 1999) or "Putnam's intuition" (Bealer, 1987; Jackson, 2011). Others describe it more expansively as the "intuition we're invited to share" (e.g., Fodor, 1987). Yet others suggest more explicit empirical claims: "most philosophers seem to share Putnam's intuitions" (Murphy, 2014), and the Twin Earth intuition is "our intuition" (e.g., Copp, 2000). Some take an even stronger line, suggesting that most ordinary people share the Twin Earth intuition—and they *ought to*: Although "some report not having" the Twin Earth intuition, "it is hard to know what to do with such people" (Korman, 2015). Importantly, while some of these descriptions are best understood as making some type of non-empirical claim, others are best understood as making straightforwardly empirical claims about people's actual intuitions and the role that those intuitions play in the philosophy of language. Our results have implications for philosophical research that makes these kind of empirical claims.

The most straightforward interpretation of our results is that they provide evidence for each separately views. Here again, this might be elaborated in a number of ways.

One way of elaborating an *each separately* view is by drawing on existing theories of *polysemy* (e.g., Nunberg, 1979; Ravin & Leacock, 2000). Broadly speaking, a term is polysemous if it has multiple senses that are closely related. A classic example is "book." This term can mean (a) a physical object (a hardbound collection of paper pages), but it can also mean (b) an abstract object (the associated work itself). Although polysemous terms like this have more than one meaning, the different meanings are closely connected to each other. In this way, cases of polysemy are very different from cases of homonymy in which words have multiple meanings that are entirely unrelated (as in the case of "bank" or "bark").

A striking fact about polysemy is that it is sometimes quite systematic (Nunberg, 1995). That is, one can often identify a general rule that characterizes the relationship between different senses across a whole class of words. To continue with our previous example, "book," "paper," and "composition" can all mean either a physical or abstract entity. In cases like these, if a speaker knows one of the meanings and also knows the general rule, that speaker should be able to derive the other meaning.

One natural hypothesis, then, would be that natural kind terms are polysemous. Much like "book" or "paper," the word "water" might be thought to have two meanings that are closely related. One meaning involves certain superficial properties; the other involves the deeper causal properties that

typically explain those very properties. As in many other cases, this polysemy would be systematic. Thus, the very same relationship between meanings found for the word "water" can also be found for "tiger," "gold," and numerous other terms.

Nichols et al. (2016) endorse this type of account. They present studies suggesting that natural kind terms sometimes take on causal–historical readings and other times take on descriptivist readings. To account for that pattern of results, Nichols et al. (2016) adopt a theory of natural kind semantics inspired by Lewis's (1999) work on semantic indecision. On their view, natural kind terms have multiple (related) senses, people can "readily switch back and forth between" these two senses, and context can determine which sense applies (Nichols et al., 2016, p. 164).

Another key question for *each separately* views concerns the role of conversational context in the semantics of natural kind terms. An obvious way to make sense of the context effect is to say that each natural kind term has multiple meanings and that people are drawn to different meanings in different contexts. But what does this show about the semantics of natural kind terms? What does the fact that people have different intuitions about a sentence in different contexts tell us about the actual truth-value of that sentence when asserted in different contexts?

One possible view would be that context actually determines a natural kind term's meaning on any given occasion of use. Thus, it might be that the word "water" has one meaning when used in scientific contexts but another meaning when used in social contexts. Another possible view would be that conversational context does not actually play a role in the semantics of natural kind terms themselves but simply has an effect on which meaning people tend to consider. On this second view, conversational context does not actually determine the meaning of the term "water." Rather, the term has two different meanings in all contexts, and it is just a fact about human psychology that people tend to consider one of these meanings in scientific contexts and a different one in social contexts. Here again, the best way of making progress on these questions will probably be to turn to existing research on broader questions in semantics, including not only theories about polysemy but also theories about the role of conversational context in semantics more generally (Preyer & Peter, 2005; Sperber & Wilson, 2012; Travis, 1981).

Insofar as empirical facts about natural kind usage provide evidence about natural kind semantics, our results most straightforwardly support *each separately* views. However, for reasons discussed in Section 6.1, one might also think that the results are compatible with certain *just causal properties* views. Specifically, one might adopt an interpretation of participants' responses according to which they are saying that the liquid in question does not count as "water" in any literal sense but can still be called "water" in a non-literal sense. This interpretation has an interesting implication regarding the context effects. Insofar as a *just causal properties* view is true, our results suggest that in social contexts we are more inclined to use some non-literal sense of "water." Intriguingly, this view entails that in many situations (e.g., social ones) our use of natural kind terms does not track the meaning of those terms.

Alternatively, it might be suggested that the responses given by our experimental participants are simply *wrong*. One might motivate this move by considering other terms besides those that we studied. For example, perhaps ordinary judgments provide good evidence about the meaning of a term like "the," but do not provide compelling evidence about the meaning of a term like "ribosome." It could then be claimed that terms like "water" are more like terms like "ribosome" than terms like "the." Like ordinary use of "ribosome," ordinary use of "water" provides poor evidence about the term's meaning.

This objection can be spelled out in a number of ways. One would be that "water" is a particularly *complex* term. Its real meaning reflects that complexity, but ordinary usage does not. Another would

be that "water" is a term that merits *deference.* There is some particular community that has privileged access to the semantics of "water," and we should defer to that group. In either case, the objection asserts that ordinary people's use of the terms is bad evidence for semantics.

The most extreme version of this view might posit that only ultra-experts have access to the relevant semantic complexity or merit deference. For instance, to understand the semantics of a term like "tiger," perhaps we need to survey PhDs in zoology. A less extreme version posits that a sufficient amount of education provides the training necessary for usage to provide reasonable evidence of semantics. For example, although online participants cannot be relied upon, an advanced PhD student's linguistic usage should provide sufficiently reliable evidence about the meaning of natural kind terms.

Our final experiment engages with this less extreme version. We looked to the natural kind judgments of elite university graduate students. The results indicate similarity in judgment among this population and the online population, providing evidence against a view that predicts elite versus non-elite natural kind intuitions differ. Moreover, both of those populations display a dual-character pattern of judgment. Insofar as we defer to "expert" concept use as evidence of meaning, our data suggest that those judgments also display a similar pattern.

In sum, although much debate remains about these semantics questions, our data most plausibly provide at least some evidence for each separately views—and against the other types of theories, including *just causal properties.*

7 | CONCLUSION

The Twin Earth thought experiment has shaped modern debates about natural kinds. By distinguishing an entity's superficial properties from its deeper causal properties, it led to the core insight that natural kinds are associated with two different representations: A set of superficial properties (e.g., a liquid's color or smell) and a set of deeper, causal properties (e.g., a liquid's underlying chemical structure).

The present studies suggest that people's ordinary judgments do not conform to the standard philosophical intuition that the deeper causal properties are the *sole* criterion of category membership. Instead, we find that people's actual judgments display a more complex dual-character pattern. Entities are categorized into natural kinds according to two different criteria. According to one, the Twin Earth liquid is water, but according to the other, it is not water.

Given our experimental findings, one might wonder why many ever thought that the "standard philosophical intuition" was widely shared. Why did it seem to so many people that these Twin Earth examples are cases in which the entity is simply not a category member (e.g., why did it seem Twin Earth liquid is not water)?

One possible answer is that there was never good reason for sharing the Twin Earth intuition in the first place. Perhaps the prevalence of a "standard" Twin Earth intuition is the result of academic sociology. As Cummins puts it:

> It is a commonplace for researchers in the Theory of Content to proceed as if the relevant intuitions [about Twin Earth] were undisputed … The Putnamian take on these cases is widely enough shared to allow for a range of thriving intramural sports among believers. Those who do not share the intuitions are simply not invited to the games. (Cummins, 1998, p. 116)

There may be some truth to Cummins's accusation, but we suspect that there is an additional factor at play. As we noted above, people's intuitions about Twin Earth cases appear to vary depending on features of the context. Now, philosophers have traditionally evaluated this case in one particular context, namely, the *philosophical* context. It is indeed possible that the intuitive response within the philosophical context is that the entity is not water, while the results reported here indicate that this is not the intuitive response in certain other contexts.

In any case, we do not regard our results as calling into question the relevance of Twin Earth cases. To the contrary, our results further support the unique significance of Twin Earth as a thought experiment that neatly distinguishes two important aspects of natural kind concepts: Superficial and causal properties. It would be a mistake, however, to assume that the standard philosophical intuition represents an empirical fact that should serve as evidence for cognitive or semantic theories. Instead, we suggest, these theories should address people's actual judgment: The liquid on Twin Earth is and is not water.

ACKNOWLEDGEMENTS

For helpful feedback, we thank Mario Attie, Pam Corcoran, Jennifer Daigle, Joanna Demaree-Cotton, Michael Della Rocca, Jussi Haukioja, Frank Keil, Clayton Littlejohn, Eddy Nahmias, Jeesoo Nam, Shaun Nichols, Daniel Weiskopf, the MindsOnline conference, and the anonymous reviewers at *Mind and Language*.

ORCID

Kevin P. Tobia https://orcid.org/0000-0003-3447-9825
George E. Newman https://orcid.org/0000-0003-0498-6746
Joshua Knobe https://orcid.org/0000-0003-0733-3775

REFERENCES

Abbott, B. (1997). A note on the nature of "water". *Mind*, *106*(422), 311–319.

Ahn, W., Kalish, C., Gelman, A., Medin, D. L., Luhmann, C., Atran, S., … Shafto, P. (2001). Why essences are essential to the psychology of concepts. *Cognition*, *82*, 59–69.

Armstrong, S. L., Gleitman, L. R. & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Bealer, G. (1987). The philosophical limits of scientific essentialism. *Philosophical Perspectives*, *1*, 289–365.

Bloom, P. (2007). Water as an artifact kind. In E. Margolis & S. Laurence (Eds.), *Creations of the mind: Theories of artifacts and their representation*. New York, NY: Oxford University Press.

Braisby, N., Franks, B. & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition*, *59*(3), 247–274.

Carey, S. (1985). *Conceptual development in childhood*. Cambridge, MA: MIT Press.

Chalmers, D. J. (2002). On sense and intension. *Noûs*, *36*, 135–182.

Cimpian, A. & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition*, *113*(1), 14–25.

Copp, D. (2000). Milk, honey, and the good life. *Synthese*, *124*, 113–137.

Corcoran, P. (2018). *The folk on twin earth* (Unpublished manuscript).

Cummins, R. (1998). Reflection on reflective equilibrium. In M. DePaul & W. Ramsey (Eds.), *Rethinking intuition*. Lanham, MD: Rowman and Littlefield.

Deutsch, M. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method*. Cambridge: MIT Press.

Fodor, J. (1987). Individualism and supervenience. In M. Boden (Ed.), *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge: MIT Press.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York, NY: Oxford University Press.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–440.

Gelman, S. A. & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*(6), 1532–1541.

Gelman, S. A. & Waxman, S. R. (2007). Looking beyond looks: Comments on Sloutsky, Kloos, and Fisher (2007). *Psychological Science*, *18*(6), 554–555.

Gelman, S. A. & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*(3), 213–244.

Genone, J. & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, *25*(5), 717–742.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, *65*(2–3), 137–165.

Horst, S. (2016). *Cognitive pluralism*. Cambridge: MIT Press.

Jackson, F. (2003). Narrow content and representation, or twin earth revisited. *Proceedings and Addresses of the American Philosophical Association*, *77*(2), 55–70.

Jackson, F. (2011). On Gettier holdouts. *Mind and Language*, *26*(4), 468–481.

Jylkka, J., Railo, H. & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology*, *22*, 37–60.

Kalish, C. W. & Gelman, S. A. (1992). On wooden pillows: Multiple classifications and children's category-based inductions. *Child Development*, *63*(6), 1536–1557.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Keil, F. C. (1995). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Symposia of the Fyssen Foundation. Causal cognition: A multidisciplinary debate* (pp. 234–267). New York, NY: Clarendon Press.

Keil, F. & Wilson, R. A. (2000). The concept concept: The wayward path of cognitive science. *Mind and Language*, *15*, 308–318.

Knobe, J., Prasada, S. & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242–257.

Korman, D. Z. (2015). *Objects: Nothing out of the ordinary*. New York, NY: Oxford University Press.

Kripke, S. (1972/1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Laurence, S., Margolis, E. & Dawson, A. (1999). Moral realism and twin earth. *Facta Philosophica*, *1*, 135–165.

Lewis, D. (1999). Many, but almost one. In *Papers in metaphysics and epistemology* (pp. 23–44). Cambridge: Cambridge University Press.

Machery, E. (2009). *Doing without concepts*. New York, NY: Oxford University Press.

Machery, E. & Seppälä, S. (2011). Against hybrid theories of concepts. *Anthropology and Philosophy*, *10*, 99–126.

Malt, B. C. (1994). Water is not H2O. *Cognitive Psychology*, *27*(1), 41–70.

Margolis, E. (1998). How to aquire a concept. *Mind and Language*, *13*, 347–369.

Medin, D. & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York, NY: Cambridge University Press.

Miller, B., & Nahmias, E. (2018). (Unpublished data).

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Murphy, L. (2014). *What makes law: An introduction to philosophy of law*. New York, NY: Cambridge University Press.

Newman, G. E. & Keil, F. C. (2008). Where is the essence? Developmental shifts in children's beliefs about internal features. *Child Development*, *79*(5), 1344–1356.

Newman, G. E. & Knobe, J. (2018). The essence of essentialism. *Mind and Language*. https://doi.org/10.1111/mila.12226.

Nichols, S., Pinillos, N. A. & Mallon, R. (2016). Ambiguous reference. *Mind*, *125*(497), 145–175.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, *3*, 143–184.

Nunberg, G. (1995). Transfers of Meaning. *Journal of Semantics*, *12*(2): 109–132.
Opfer, J. E. & Siegler, R. S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology*, *49*(4), 301–332.
Pessin, A. & Goldberg, S. (1996). *The twin earth chronicles: Twenty years of reflection on Hilary Putnam's the meaning of meaning*. New York, NY: Routledge.
Preyer, G. & Peter, G. (2005). *Contextualism in philosophy*. Oxford: Oxford University Press.
Putnam, H. (1973). Meaning and reference. *Journal of Philosophy*, *70*(19), 699–711.
Putnam, H. (1975). The meaning of meaning. In *Minnesota studies in the philosophy of science: Language, mind, and knowledge* (Vol. 7, pp. 131–193). Minneapolis: University of Minnesota Press.
Ravin, Y. & Leacock, C. (2000). *Polysemy: Theoretical and computational approaches*. Oxford: Oxford University Press.
Sassoon, G. W. (2013). A typology of multidimensional adjectives. *Journal of Semantics*, *30*(3), 335–380.
Sperber, D. & Wilson, D. (2012). *Meaning and relevance*. New York: Cambridge University Press.
Strevens, M. (2000). The essentialist aspect of naïve theories. *Cognition*, *74*(2), 149–175.
Travis, C. (1981). *The true and the false: The domain of the pragmatic*. Philadelphia, PA: John Benjamins Publishing Company.
Weiskopf, D. A. (2009). The plurality of concepts. *Synthese*, *169*, 145–173.

SUPPORTING INFORMATION

Additional supporting information may be found online at https://osf.io/f44hq/.

How to cite this article: Tobia KP, Newman GE, Knobe J. Water is and is not H_2O. *Mind Lang*. 2019;1–26. https://doi.org/10.1111/mila.12234