

Modal Intuitions Going Two-Dimensional

forthcoming in *Oxford Studies in Experimental Philosophy* vol. 6

Adrian Ziółkowski¹, Krzysztof Śekowski², Anand Jayprakash Vaidya³

¹ Faculty of Philosophy, University of Warsaw, adrian.a.ziolkowski@uw.edu.pl

² Faculty of Philosophy, University of Warsaw, k.sekowski@uw.edu.pl

³ Department of Philosophy, San Jose State University

Abstract

In the contemporary study of modality, two important theories diverge in their analysis of claims about the essential properties of natural kinds. The one-dimensional metaphysical approach asserts that there is no interpretation under which claims such as ‘water could have had the chemical structure XYZ’ are true, whereas two-dimensional epistemic semantics allows for such an interpretation. The paper explores how empirical studies of folk intuitions can contribute to the debate between these views. We investigated whether laypersons can adopt a two-dimensional epistemic perspective when evaluating modal claims involving natural kind terms. We conducted two studies to determine if lay speakers can conceive of possible worlds where a natural kind term such as ‘water’ refers to entities with different essential features (e.g., water is XYZ) and interpret modal claims in accordance with two-dimensional semantics. The findings indicate that folk semantic judgments tentatively support the predictions of two-dimensional semantics. Moreover, our study provides valuable methodological insights, highlighting the influence of familiarity bias introduced by specific terms, such as ‘water,’ which underscores the need for caution in the design of study materials in experimental semantics.

Introduction

In the contemporary study of the philosophy of modality, there are two important approaches to the semantics of modal claims, such as ‘water could have had the chemical structure XYZ.’ The one-dimensional metaphysical approach, pioneered by Kripke (1980), holds that there is no interpretation under which the sentence ‘Water could have had the chemical structure XYZ (even though it actually is H₂O)’ is true. The two-dimensional epistemic¹ approach, pioneered by Chalmers (1996, 1999, 2002, 2004), holds that there is an interpretation under which it is true.

How can we determine which view of modal semantics is more plausible? There are at least two kinds of considerations that can be brought to bear. On the one hand, there are theoretical considerations from the philosophy of language, logic,

¹ Chalmers (2004) uses the phrase ‘epistemic two-dimensional semantics,’ or ‘E2D.’ For the purposes of contrasting his approach with Kripke’s approach which is metaphysical in nature; we have placed ‘epistemic’ after ‘two-dimensional’.

metaphysics, and epistemology. On the other hand, there are empirical considerations deriving from studies of folk intuitions about modal claims.

There is a rich literature in the philosophy of language concerning the plausibility of two-dimensional semantics (e.g., Chalmers 2004, Garcia-Carpintero & Macias 2004, Soames 2005).² And there is some literature on the background logic, S5, that supports two-dimensionalism (Salmon 1989, Vaidya 2008). However, despite a rich literature on various types of semantic intuitions, their consistency across demographic and other factors, and their coherence with a variety of semantic theories (Genone & Lombrozo 2012; Machery et al. 2004; Machery 2017; van Dongen et al. 2021), there is almost no data that inform us about intuitions concerning modal claims. There are some notable exceptions, such as the extensive literature on counterfactuals, which is mostly related to discussions regarding causality (e.g., Icard et al. 2017, Quillien 2020), and empirical studies concerning counternormative statements and the phenomenon of imaginative resistance (e.g., Liao et al. 2014, Barnes & Black 2016); however, none of these research fields investigates modal semantic judgments *per se* (see also Knobe & Szabó (2013) on possible hidden modality in lay speakers' interpretations of certain linguistic constructions).

In this paper, we empirically investigate semantic intuitions about modal claims, such as 'water could have had the chemical structure XYZ,' and check whether lay speakers are capable of adopting the two-dimensional epistemic perspective when considering modal properties of linguistic expressions.

Based on the data collected in two studies, we will argue that lay speakers are capable of: (1) conceiving of possible worlds in which a given kind term (e.g., 'water') refers to objects possessing different essential features than the essential features shared by what is picked out by this term in the actual world (e.g., imagining a possible world where water is XYZ), and (2) evaluating modal claims along the lines of two-dimensional semantics, taking such possible worlds as evaluation points for their truth-value judgments regarding modal claims. We believe that our results provide tentative support for the claim that the predictions of two-dimensional semantics are reflected in folk semantic judgments regarding modal utterances.

In Section 1, we briefly outline the differences between one-dimensional and two-dimensional possible world semantics. Section 2 summarizes the results of pre-existing empirical research concerning issues relevant to the topic of our endeavor, such as the reference of natural kind terms, psychological essentialism, and

² 'Two-dimensional semantics' is an umbrella term and there are many accounts that are classified as representative of this approach, most notably those discussed in the seminal work of Stalnaker (1978) and Kaplan (1989). However, in this paper we focus on the view known as epistemic two-dimensional semantics, articulated and defended most explicitly by Chalmers in his (2004).

semantic externalism. We also explain why our proposal is a novel approach to the issue at hand. In Section 3, we explain the motivation for our study, present our hypotheses, provide a detailed description of the methods employed in our first experiment, and report the results. Section 4 offers two alternative interpretations of the data collected in Experiment 1 and explains the rationale for conducting a follow-up study. Section 5 details the methods used in the follow-up and reports its outcome. Finally, in Section 6, we summarize our findings and present what we believe is the upshot of our study.

1. One-dimensional Metaphysical Semantics vs. Two-dimensional Epistemic Semantics

In post-Kripkean studies of the epistemology of modality, a well-known problem arises for *a priori* accounts of the relation between conceivability and possibility. The problem is articulated in the work of Yablo (1993), we call it the *hostage to a posteriori necessity problem* for modal rationalism—the view that metaphysical possibility is knowable *a priori*.

According to Kripke (1980), there are *a posteriori* necessities—necessary truths that are only knowable *a posteriori*. For example, Kripke offers the following deduction model of how we know that it is necessary that water = H₂O.

1. Where x and y are variables ranging over rigid designators, if $x = y$, then necessarily $x = y$.
2. 'Water' and 'H₂O' are rigid designators.
3. Water = H₂O.
4. So, it is necessary that water = H₂O.

Kripke argues that given that (3) is only knowable *a posteriori*, no matter whether (2) and (1) are knowable *a priori*, (4) is known *a posteriori*.

Using Kripke's conception of *a posteriori necessities* and how we know them, Yablo (1993) draws out the hostage problem for *a priori* approaches to the relation between conceivability and possibility.

1. Suppose it is *prima facie* conceivable that q .
2. Suppose it is necessary that p , but only knowable *a posteriori*, such as on a Kripkean model of knowledge of empirical necessities.
3. Suppose that if p , then not- q .
4. It follows that even though q is *prima facie* conceivable, one could never know *a priori* that it is not possible that q because the necessity of p is only knowable *a posteriori*.

Chalmers (2002) uses an epistemic interpretation of two-dimensional semantics, E2D, to circumvent the hostage problem. The debate between him and Kripke can be seen in their contrasting answers to the question: could water have been XYZ?

Kripke claims that given that water is H₂O in the actual world, it could not have been the case that water is XYZ. Consequently, any subject that claims to have conceived of water without H₂O has failed to conceive of anything. Chalmers argues that given that water could have turned out to be XYZ, it could have been XYZ. To get this result, Chalmers uses E2D.

In contrast to one-dimensional semantics, two-dimensional semantics assigns extensions and truth-values to expressions relative to two possible world parameters, rather than just one. One important way to see the difference between one- and two-dimensional semantics is from a formal model-theoretic point of view. On one-dimensional semantics, there is a model $\langle W, R, V \rangle$ where W is the set of possible worlds, R is an accessibility relation on W , and V is a valuation function that delivers for any proposition p specified at a world w_1 , what its truth value is at another world w_2 if w_2 is accessible from w_1 . In addition, Kripke claims that there will be a distinguished member of W , g , which is intuitively the actual world. Two-dimensional semantics has a similar model; however, there is no distinguished member, g . Any world can be taken to be the actual world. Although there are various interpretations of two-dimensional semantics, for capturing the debate between Kripke and Chalmers over whether water could have been XYZ, we will simplify.

In one-dimensional semantics, the standard question is: given that g is the actual world, and 'T' refers to O in g , what does 'T' refer to in another possible world, say w_1 ? For example, given that 'water' refers to H₂O in the actual world, what does 'water' refer to in w_1 , where there is no H₂O, only XYZ? Kripke answers that it refers to nothing in w_1 .

By contrast, in Chalmers's two-dimensional epistemic semantics, the absence of a distinguished actual world allows for one to ask two distinct questions associated with distinct intensions. First, in line with Kripke, there is the counterfactual question, associated with the secondary intension: suppose w_1 is the actual world, where 'T' refers to O, what does 'T' refer to in w_2 ? Second, departing from Kripke, there is the actual world question, associated with the primary intension: suppose w_1 turns out to be the actual world, what does 'T' refer to in w_1 ? Chalmers argues that while the counterfactual question in two-dimensional semantics is the same as the standard question in one-dimensional semantics, the actual world question is unavailable in the one-dimensional framework. He further argues that we can get to the conclusion that water refers to XYZ if we consider w_1 to be the actual world, by taking the description associated with water in g , call it D, and seeing what in w_1 is picked out by D. Let D = 'the local potable liquid that falls from the skies, quenches

thirst, covers 70 percent of the Earth, and flows in rivers and streams, etc.’ Chalmers reasons that if we take w_1 to be a hypothesis about the actual world, namely that it is the actual world, and it turns out that in w_1 D picks out XYZ, and not H₂O, then we are rationally led to the conclusion that water = XYZ in w_1 . That is: it could turn out that water = XYZ. On Chalmers’ account, the primary intension of ‘water’ leads to an analysis where it is primarily possible that water = XYZ, and an account of the secondary intension of ‘water’ where it is secondary impossible that water = XYZ. The difference between these two types of intensions affects the truth values of sentences depending on the world in which they are expressed, which, together with Chalmers’ modal monism, opens a way to state that water could have been XYZ (Chalmers 2002), while for Kripke, this is simply impossible.

Which of the two rival approaches to modal semantics, one-dimensional or two-dimensional, is better reflected in lay speakers’ modal intuitions? In this paper, we empirically address this question.

2. A brief look at the pre-existing empirical data

Before turning to the study, let us review the state of the art in experimental philosophy concerning related areas of research. We will discuss studies on reference, psychological essentialism, and the meaning of natural kind terms, aiming to show that their results partly support the hypothesis that lay speakers might interpret the meaning of natural kind terms in two ways, as two-dimensional semantics suggests. The general picture emerging from studies on folk semantic intuitions is that, whereas sometimes the meaning of an expression is determined by external properties such as the purported essential properties of the referent (see e.g., Nichols et al. 2016, Tobia et al. 2020), on other occasions it is, at least partly determined by our identificatory knowledge of the object (see Machery & Seppälä 2011, Genone & Lambrozo 2012, Knobe et al. 2013, Haukioja et al. 2021). These results justify our expectation that semantic intuitions might align with the predictions of two-dimensional semantics.

The first group of results that modestly support the expectation that two-dimensional semantics might be reflected in semantic intuitions comes from studies on the reference of proper names. Early experimental studies on the theory of reference (Machery et al. 2004) found cross-cultural differences in semantic intuitions, showing that people from Western cultures lean towards the causal theory of reference more than people from Eastern cultures. As meta-analyses of experimental philosophy of language suggest, cross-cultural differences in this respect are supported by a variety of studies (Machery 2017, van Dongen et al. 2021). However, it is controversial how to interpret such data. Importantly, besides cross-cultural differences, these studies also observed a huge variability of intuitions

within each cultural group. As Knobe argues, those results might be interpreted not as evidence that people systematically differ in the intuitions they hold, but rather as showing that it is common for people to have *internally* conflicting intuitions (Knobe, forthcoming). He also points out the stability of patterns regarding differences in verdicts about thought experiments (see also Knobe 2021, 2023).

Some suggested that the mentioned data support the view that there are several theories of reference rather than a single theory of reference (Machery et al. 2004). In such a case, these results would not align with the predictions of two-dimensional semantics, which posit that the process of determining the meaning or reference of expressions should be universal but can be interpreted in two ways, following primary or secondary intension. Two-dimensional semantics account for both externalist (e.g., causal) and epistemological (e.g., identificatory knowledge) factors in establishing reference, suggesting either can play a role depending on the kind of intension following which the claim is interpreted.

Later studies in experimental philosophy of language have provided results even more promising for two-dimensional semantics. These studies have shown that semantic intuitions are influenced by both types of factors. For instance, studies that report evidence in favor of hybrid theories of reference (e.g., Genone & Lombrozo 2012) and studies indicating that terms can be contextually ambiguous, influenced by either epistemological or external factors (Nichols et al. 2016), support this view.

Similar results are found in studies on natural kind terms, particularly those on psychological essentialism. Psychological essentialism states that people categorize objects based on the belief that they possess an essential characteristic of their category. This can be understood in two ways:

- 1. Externalist essentialism:** Speakers believe that X falls under a natural kind concept C because X has an essential property E.
- 2. Hybrid essentialism:** Speakers believe that X falls under a natural kind concept C in one sense (externalist), because X has an essential property E, and in another sense (epistemic) because it fits the identificatory knowledge of C.

In a study by Jylkkä et al. (2009), participants were presented with a story about a mineral, *zircaum*, found in Siberia. The story continued with scientists later discovering another mineral with the same external properties (such as color, smell, and taste) in Norway, and also naming it 'zircaum.' However, a few weeks later, scientists determined that the mineral found in Norway had a different chemical structure from the one found in Siberia. Participants were asked whether the substance found in Norway was zircaum twice: before being informed about the discovery of its chemical structure and after being presented with this information.

Additionally, at the end, they were asked whether their initial answer about the mineral was true or false, given the later discovery. While most responses aligned with externalist essentialism, a significant number seemed to be shaped by epistemological factors. Similar mixed results, showing the importance of both internalist and externalist factors, were found in further studies employing Twin-Earth-like scenarios (e.g., Tobia et al. 2020; Haukioja et al. 2021) or other, similar scenarios (Machery & Seppälä 2011, Knobe et al. 2013).

The mixed nature of these results suggests that either, at least in some contexts or senses, expressions are interpreted in a purely externalist way, while in others, in a more internalist manner, or that both internalist and externalist factors simultaneously work as necessary conditions for falling within the extension of natural kind terms. However, what is crucial for our study is that none of the discussed experiments addressed modal claims. Thus, while these results provide modest support for the hypothesis that folk semantic intuitions might align with the predictions of two-dimensional semantics, they do not offer a definitive answer. Hence, our study aims to provide new evidence that could bring us closer to a definitive answer. We treat the results briefly described above as a rationale to posit a positive hypothesis on the relation between semantic intuitions and two-dimensional semantics.

3. Experiment 1

The main goal of the study was to test whether lay speakers are capable of (1) conceiving of possible worlds in which a given kind term (e.g., ‘water’) refers to objects possessing different essential features than the essential features shared by what is picked out by this term in the actual world (e.g., imagine a possible world where water is XYZ), and (2) evaluate modal claims along the lines of two-dimensional semantics, taking such possible worlds as evaluation points for their truth-value judgments regarding modal claims (e.g., ‘Although water is XYZ, it could have a different chemical structure, and be H₂O’).

Let us clarify what the goal of our study is, and what it is not. We do not aim to provide evidence for preferring two-dimensional over one-dimensional semantics. Although such a claim would be highly desirable, it is probably too ambitious given, on the one hand, that these theories may yield overlapping predictions concerning many modal claims, because both preserve a Kripkean treatment of necessity at the level of secondary intension, and, on the other hand, that a defender of one-dimensionalism may explain judgments that (purportedly) support the predictions of two-dimensional semantics by appealing to some type of semantic confusion (Kripke 1980: 128-129; 160-161). We do aim, however, to test whether lay

modal intuitions support predictions made by two-dimensional semantics and to examine whether these intuitions form a stable pattern.

Regarding the coherence with the predictions of two-dimensional semantics, the main idea is to determine whether people would judge the possibility of a certain natural kind lacking its essential property by distinguishing between primary and secondary intensions. For example, consider the term 'water' and the essential property of its referent, which is its chemical structure. We expect that most people would say that it is impossible (in the secondary sense of possibility) for water to have a different structure than it essentially does in a world considered as actual. Importantly, people should not just say it is impossible for water to lack the chemical structure H_2O relative to our actual world. If their judgments are to be in line with two-dimensional semantics, they should also recognize it as impossible for water to have a different chemical structure relative to any possible world, where the term 'water' refers to something other than H_2O (e.g., XYZ), following the primary intension of the term. So, for instance, if in the world under scrutiny 'water' refers to XYZ, people should be able to conceive of such a world, and also recognize the impossibility of water having a chemical structure different from XYZ (in the secondary sense).

Our aim is therefore not simply to ask whether people think that water might have had a different structure from the one it actually has. In such a design, regardless of whether participants accept or deny claims such as 'Water might have had the chemical structure XYZ,' it would remain unclear whether their judgments were driven by a primary or a secondary interpretation. To avoid this ambiguity, our experiment involved two connected parts. First, participants were asked to treat a stipulated world as actual and determine what the relevant natural-kind term referred to in that world. Second, they were asked to evaluate a counterfactual modal claim relative to that world by assuming they inhabit this world and they are the addressees of the claim. The design therefore allowed us to examine whether participants could switch from a primary-intensional perspective, in which the stipulated world is considered as actual, to a secondary-intensional evaluation of counterfactual possibilities relative to that world.

If we observe a pattern of responses showing that participants can treat different possible worlds as actual, determine the extension of the relevant term relative to each of them, and then preserve that locally fixed extension across counterfactual evaluation, it would not by itself favor two-dimensional over one-dimensional semantics. Recall that an opponent of two-dimensionalism could still try to explain away such results by providing an error theory of relevant judgments. However, the broader and more stable the observed pattern, the wider the range of judgments that a defender of pure one-dimensionalism would need to explain as resulting

from confusion or some other source of error, which would put more argumentative burden on the opponents of two-dimensional semantics.

The latter point brings us to the question of the stability of intuitions. The existing experimental philosophy literature shows that intuitions may be highly sensitive to philosophically irrelevant factors, such as interpersonal differences (e.g., culture or gender), and framing effects (see e.g., Machery et al. 2004; Machery 2017; van Dongen et al. 2021; Sękowski et al. 2023; Stich & Machery 2023; but for arguments for the stability of intuitions across demographic factors, see e.g., Knobe 2021, 2023). Our approach is to test not only whether intuitions align with two-dimensional semantics as described but also whether this coherence is stable. In other words, we seek to determine empirically whether results concerning modal intuitions can be theoretically useful for modal semantics. To this end, we compare the similarity of responses across scenarios involving different kind terms and across worlds that differ in their similarity to the actual world.

Summing up, we designed our experiment not only to see whether we can obtain data that support positive answers to questions (1) and (2) posed at the beginning of this section, but also to explore avenues where laypersons' intuitions might conflict with the predictions of two-dimensional semantics. The next section provides a detailed description of our methods and materials used in the study.

3.1. Methods: participants, materials, and procedure

The experiment employed two scenarios involving kind terms: *Water*, and *Cats* (concerning the kind WATER and CAT, respectively). The participants were asked to evaluate modal claims regarding the kind in question, which were either uttered in a possible world *Similar* to the actual world (e.g., where water is H₂O) or its *Dissimilar* counterpart (e.g., where water is XYZ). Each participant was randomly assigned to one *Scenario* and one *World* condition, which yields a 2×2 between-subjects design.

The study was conducted online. After agreeing to participate in the study, respondents were presented with a short introduction; for the sake of illustration, we provide the contents of the introduction that preceded the *Water* scenario (in both *Similar* and *Dissimilar* conditions).

Here are two different kinds of colorless, tasteless liquids that can quench thirst and support life but differ with respect to their chemical structure:

- a) H₂O [a picture of a glass of water]
- b) XYZ [the same picture as in (a)]

H₂O and XYZ look and taste exactly the same.

Afterwards, participants read the scenario in one of the two variants:

Water: Similar/Dissimilar

Imagine a world that is pretty much like ours. [What]_{similar} / [However, the history of this world took a different turn, and, as a result, what]_{dissimilar} can be found in the oceans, lakes and rivers, what is approximately 60% of a human's body, has the structure [H₂O]_{similar} / [XYZ]_{dissimilar} and is called water [, just as is the case in our world.]_{similar} / [.]_{dissimilar} All water is exactly like that. There is no fluid of the structure [XYZ]_{similar} / [H₂O]_{dissimilar} in this world.

Right after reading the scenario, subjects were asked to complete the following: (C) 'Suppose that you live in this world. Water is: [H₂O; XYZ].' Task (C) served both as a comprehension and manipulation check; only those subjects who chose 'H₂O' in the *Similar* and 'XYZ' in the *Dissimilar* condition were included in the final analysis.

The next section of the survey included the target measure (truth-value judgment regarding the relevant modal utterance; TVJ) and two follow-up questions:

(TVJ) Suppose that you live in this world. You meet your friend who says: 'Although water is [H₂O]_{similar} / [XYZ]_{dissimilar}, water could have a different chemical structure, and be [XYZ]_{similar} / [H₂O]_{dissimilar}.' What your friend said is: [True; False]

(*Confidence rating*) How certain are you of your previous answer? [5-point Likert scale ranging from 'Completely uncertain' to 'Completely certain']

(*Imaginability rating*) How easy is it for you to imagine such a world? [5-point Likert scale ranging from 'Very difficult' to 'Very easy']

When the participants answered all these questions and proceeded to the next section, they were presented with the attention check—an instruction that asked them to select 'No answer' out of three answer options ('Yes'; 'No'; 'No answer') and type 'I am attentive' in the indicated comment box. Only those respondents who fulfilled both requirements were included in the final analysis. The final section of the survey collected basic demographic data concerning the participants (their age, gender, and education).

The procedure for the *Cats* scenario was analogous. The introduction provided a brief description of two types of entities that look and behave similarly, but one are biological creatures that belong to the group of mammals (cats), and the other are artificial beings, composed of synthetic parts (robotic cats). Each type was illustrated with the same picture. The vignette, depending on the *World* condition, was as follows:

Cats: Similar/Dissimilar

Imagine a world that is pretty much like ours. [Entities]_{similar} / [However, the history of this world took a different turn, and as a result, entities]_{dissimilar} that spend their time catching mice, birds, sleeping, meowing, and purring are [mammals]_{similar} / [robots]_{dissimilar} and are called cats[, just as is the case in our world.]._{similar} / [.]_{dissimilar} All cats are exactly like this. There are no entities that behave like this and are [robots]_{similar} / [mammals]_{dissimilar} in this world.

After answering the control question, which was formulated in an analogous way as in the *Water* case, participants were asked to evaluate the target modal utterance: ‘Although cats are [mammals]_{similar} / [robots]_{dissimilar}, cats could have a different internal structure, and be [robots]_{similar} / [mammals]_{dissimilar}.’ The main study question (Truth-value judgment) was presented alongside the Confidence and Imaginability rating questions, similarly to the *Water* scenario. The same attention check was employed, and the survey concluded with the demographics section.

The main dependent variable in the study was the dichotomous truth-value judgment regarding the target modal claim. Additionally, we combined subjects’ truth-value judgments (‘True’ coded as ‘1’; ‘False’ coded as ‘-1’) with their responses to the confidence question (ranging from ‘1’—‘Completely uncertain’ to ‘5’—‘Completely certain’) by multiplying their numerical values. The resulting composite score was used in an additional analysis that could potentially detect more subtle differences in participants’ modal intuitions across experimental conditions.

All the relevant details of the study were preregistered, and the preregistration form was published on the OSF website (<https://osf.io/2txpa>) prior to data collection.

The participants were recruited on Prolific (<https://www.prolific.com/>). For taking the survey, they received a small monetary compensation (£0.45). 347 respondents completed the survey, but 65 of them did not pass the attention check; additional 28 responses were excluded from the final analysis due to failing the comprehension check, which left us with a final sample size of $N = 254$. Of these

participants, 49.6% self-identified as female, 49.2% male, and 1.2% non-binary. Their age ranged from 18 to 77 years ($M = 41.11$; $SD = 13.94$).

3.2. Research hypotheses

Although our study aimed at finding whether lay semantic judgments are compatible with the predictions of two-dimensional semantics, we did not hypothesize a positive answer to this question. According to our main hypothesis:

(H1) Overall, respondents will be more likely to give positive than negative truth-value judgments (TVJs) when evaluating the target modal claim.

Since we designed our study materials to operationalize the distinction between primary and secondary intensions, and the target modal claim should be interpreted secondarily, two-dimensional semantics would predict that it is false. Hence, H1 bearing out would provide *prima facie* evidence against two-dimensional semantics.³

Importantly though, H1 is not the only way our results might be at odds with the predictions of two-dimensional semantics. It might also turn out that lay modal intuitions concerning matters so intricate as those described in our study materials are incoherent or confused in the first place, and are not compatible with either semantic theory discussed here. Such a conclusion could be drawn if we confirm either of the following:

(H2) Participants will provide different truth-value judgments (TVJs) depending on *Scenario* manipulation (i.e., for different kinds of entities in question).

(H3) Respondents will give different truth-value judgments (TVJs) depending on *World* manipulation (i.e., for worlds more or less similar to the actual world).

³ It is worth noting that, should our study find that participants tend to give positive TVJs, this tendency might stem from various sources. It might be due to respondents' inability to properly distinguish between primary and secondary intensions. But it could also indicate that lay modal intuitions are at odds with both Chalmers' and Kripke's accounts, and that laypersons have different beliefs on essential properties of a given natural kind members, or are not essentialists in the first place. However, this issue is irrelevant to us, as the main aim of the study is to test whether lay modal judgments can be reconciled with the two-dimensional framework.

Again, our hypotheses are negative. According to two-dimensional semantics, semantic rules governing truth conditions for modal claims regarding natural kinds should remain the same, no matter what kind or world we consider. H2 and H3 predict that our participants' responses will not display this level of consistency.

Notice that, if we do not find evidence in favor of any of our hypotheses and, moreover, we obtain evidence against H1, namely, that respondents are more likely to judge the target modal utterance false than true consistently across all the experimental conditions, our data would be in line with the predictions of two-dimensional semantics.

3.3. Experiment 1: Results

3.3.1. Results: truth-value judgments

The first thing worth emphasizing is that our data indicates that laypersons are capable of conceiving of possible worlds in which a given kind term denotes objects different in their essence from objects picked out by the term in the actual world. Out of 282 respondents who passed our attention check, only 28 (9.9%) failed the comprehension check (e.g., said that water is H₂O in a possible world where the term 'water' picks out XYZ). This means that the primary reading of statements such as 'Water could have been XYZ' is readily accessible to laypersons.

To check whether our experimental manipulation affected participants' truth-value judgments regarding the target modal utterance, we ran a logistic regression analysis on the data. The results are summarized in Table 1.

Table 1. Logistic regression for truth-value judgments in Experiment 1: main effects and interaction.

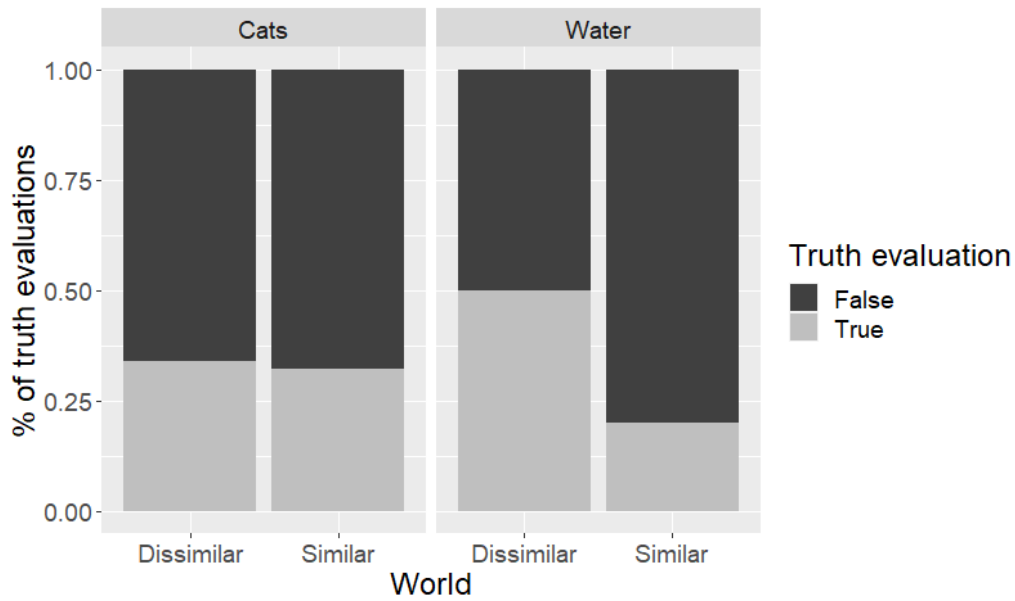
Predictor	<i>B</i>	<i>SE B</i>	<i>z-value</i>	<i>p-value</i>
Scenario				
Cats	—	—	—	—
Water	0.67	0.35	1.93	0.054
World				
Similar	—	—	—	—
Dissimilar	-0.07	0.38	-0.19	0.85
Scenario × World				
Cats – Similarity	—	—	—	—
Water – Similarity	-1.31	0.56	-2.35	0.019
Constant	-0.67	0.26	-2.62	0.009

Contrary to our second hypothesis, we did not find a main effect of *Scenario* ($p = 0.054$) or *World* ($p = 0.85$) manipulations. However, we did observe a significant interaction between these two factors ($B = -1.31$, $z = -2.35$, $p = 0.019$) which emerged because our participants were more likely to judge the target utterance true in the

Water-Dissimilar condition than in the *Water-Similar* condition ($z = 3.04$, $p = 0.002$; two-sided), whereas the *World* manipulation had no impact on subjects' truth-value judgments for the *Cats* scenario.

When it comes to the distribution of truth-value judgments in each experimental condition, in three out of four conditions, the majority of respondents judged the target modal utterance false; binomial tests, two-sided: *Cats-Similar* ($p = 0.009$), *Cats-Dissimilar* ($p = 0.010$), *Water-Similar* ($p < 0.001$). In the *Water-Dissimilar* condition, the judgments split perfectly in half and did not differ from chance distribution. The results are illustrated in Figure 1.

Fig. 1. The proportion of truth-value judgments depending on the experimental condition, Experiment 1.



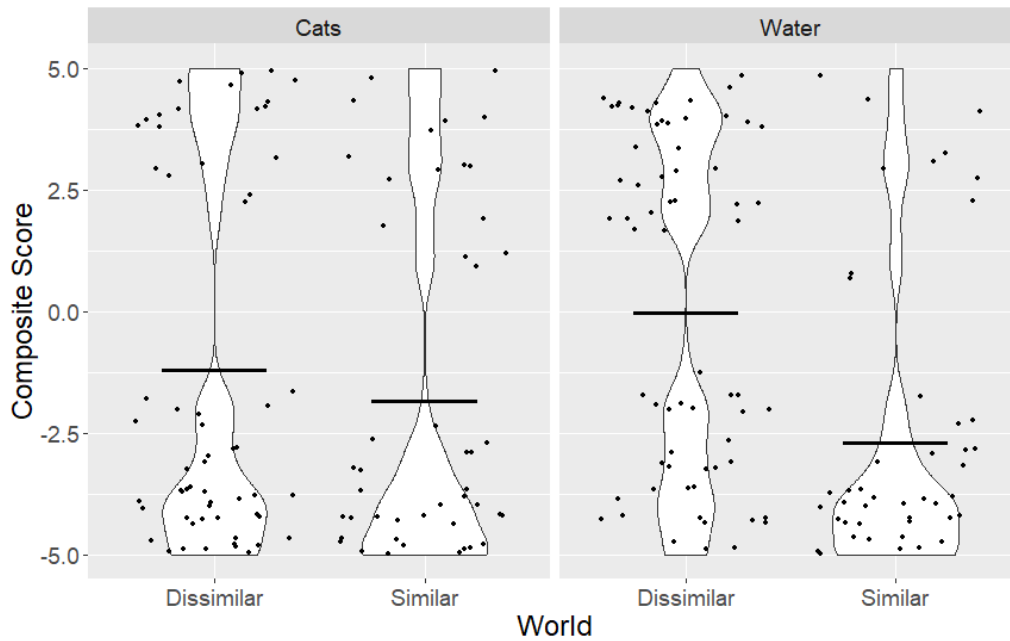
3.3.2. Results: composite scores

For an additional analysis, we used the composite score, which combines subjects' truth-value judgments with their confidence. This analysis could potentially detect more subtle differences in participants' intuitions across experimental conditions.

To analyze the data, we subjected it to a 2×2 ANOVA analysis including interactions. Although no significant main effect of *Scenario* emerged ($F(250, 1) = 0.48$; $p = 0.49$, $\eta^2 = 0.001$), we observed a significant effect of *World* manipulation ($F(250, 1) = 12.89$; $p < 0.001$, $\eta^2 = 0.049$). On average, participants' scores were higher in the *Similar* condition ($M = -0.61$; $SD = 3.79$) than in the *Dissimilar* condition ($M = -2.26$; $SD = 3.45$). Similarly as in the case of our main analysis involving dichotomous truth-value judgments, we found a significant interaction between the two factors for the composite score ($F(250, 1) = 4.99$; $p = 0.026$, $\eta^2 = 0.020$), which resulted from a stronger impact of *World* manipulation on participants' scores for the *Water* scenario than the *Cats* scenario.

To further explore the interaction, we performed a series of pairwise comparisons of the average scores across experimental conditions using Tukey HSD. Although for the *Cats* scenario there was no significant difference between the *Dissimilar* ($M = -1.22$; $SD = 3.92$) and *Similar* ($M = -1.85$; $SD = 3.74$) conditions ($p = 0.74$), an effect of *World* manipulation emerged for the *Water* scenario with higher overall scores in the *Water-Dissimilar* ($M = -0.04$; $SD = 3.60$) than in the *Water-Similar* ($M = -2.71$; $SD = 3.09$) condition ($p < 0.001$). The average scores are presented in Figure 2.

Fig. 2. The distribution of composite scores depending on the experimental condition in Experiment 1. Negative values indicate negative verdicts regarding the target modal utterance, positive values indicate the opposite. Vertical bars represent the average score.



3.3.3. Results: confidence ratings

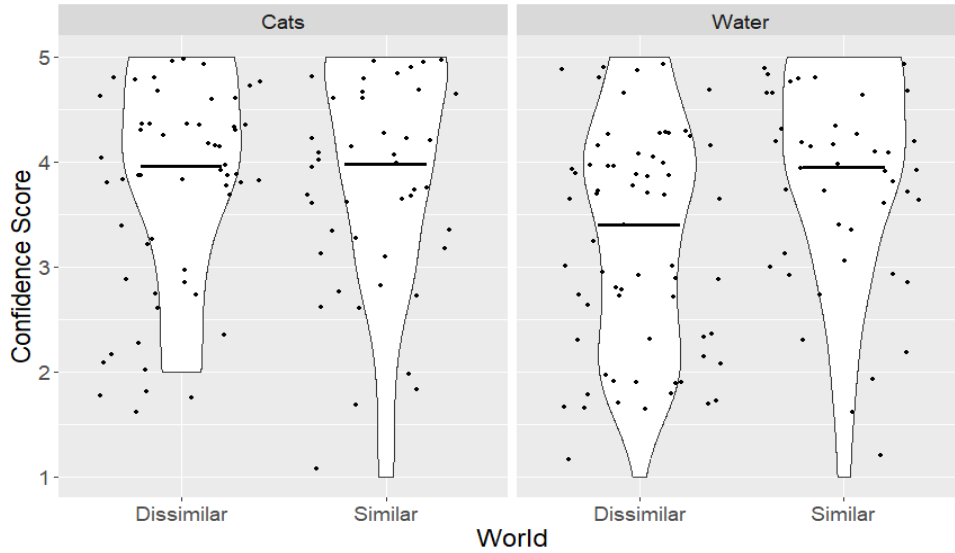
Although we did not formulate any hypotheses concerning the impact of our experimental manipulation on confidence ratings, we will briefly report the findings regarding this measure, as we believe they shed light on the main result we reported in the previous sections.

A quick look at the average confidence ratings (see Figure 3) reveals that the *Water-Dissimilar* condition is an outlier also for this measure. 2×2 ANOVA analysis showed both a significant effect of *Scenario* ($F(250, 1) = 5.95$; $p = 0.015$, $\eta^2 = 0.022$) and *World* manipulation ($F(250, 1) = 4.31$; $p = 0.039$, $\eta^2 = 0.017$), whereas the interaction between the two factors was marginally significant ($F(250, 1) = 3.57$; $p = 0.060$, $\eta^2 = 0.014$). Overall, the respondents felt more confident with their truth-value judgments when evaluating the *Cats* scenario ($M = 3.97$; $SD = 1.07$) than when judging the *Water* case ($M = 3.64$; $SD = 1.12$) and exhibited less confidence in the

Dissimilar condition ($M = 3.67$; $SD = 1.09$) than in the *Similar* condition ($M = 3.96$; $SD = 1.10$).

To further explore the differences between experimental conditions, we performed post-hoc pairwise comparisons (Tukey HSD). Although we found no differences between the *Similar* ($M = 3.98$; $SD = 1.14$) and *Dissimilar* condition ($M = 3.96$; $SD = 1.01$) for the *Cats* scenario ($p > 0.99$), for the *Water* scenario, participants exhibited significantly less confidence in their judgments in the *Dissimilar* ($M = 3.40$; $SD = 1.10$) than in the *Similar* ($M = 3.95$; $SD = 1.08$) condition ($p = 0.028$). Indeed, the two main effects were mostly driven by the drop in confidence in one particular condition, *Water-Dissimilar*, where respondents' confidence was significantly lower than in all the remaining three cases (all p 's < 0.05).

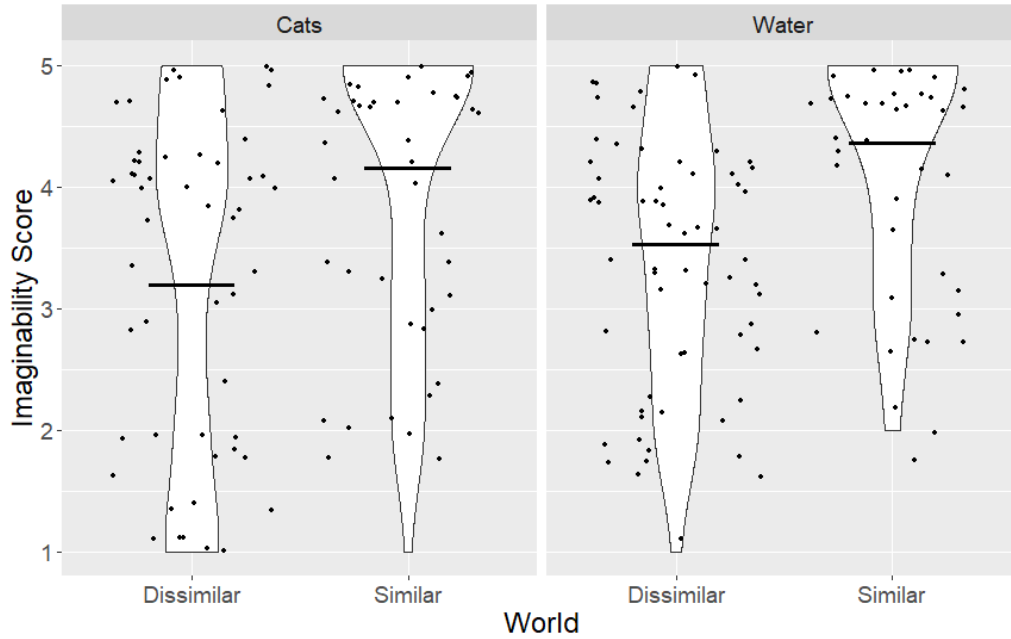
Fig. 3. The distribution of confidence ratings depending on the experimental condition in Experiment 1. '5' — 'completely certain'; '1' — 'completely uncertain'. Vertical bars represent the average score.



3.3.4. Results: imaginability ratings

When it comes to imaginability ratings, the analysis found that only the *World* factor had a significant impact on subjects' self-reported difficulties with imagining the presented scenarios ($F(250, 1) = 33.31$; $p < 0.001$, $\eta^2 = 0.118$). On average, participants found it harder to conceive of *Dissimilar* possible worlds ($M = 3.36$; $SD = 1.35$) than to imagine *Similar* worlds ($M = 4.26$; $SD = 1.09$). The magnitude of this effect was comparable for the *Cats* and *Water* scenarios, as no main effect of *Scenario* ($F(250, 1) = 2.65$; $p = 0.105$, $\eta^2 = 0.013$) or interaction ($F(250, 1) = 0.16$; $p < 0.687$, $\eta^2 < 0.001$) emerged. These results are illustrated in Figure 4.

Fig. 4. The distribution of imaginability ratings depending on the experimental condition in Experiment 1. ‘5’—‘very easy [to imagine]’; ‘1’—‘very difficult [to imagine]’. Vertical bars represent the average score.



4. Discussion

The data collected in our first study shows that, although laypersons report minor difficulties with imagining possible worlds where some kinds of objects have different essential features than in the actual world (e.g., water is XYZ), they do accept scenarios describing such worlds as conceivable. Moreover, they are capable of judging the truth values of sentences relative to such possible worlds, which suggests that the theoretical perspective of two-dimensional semantics is available to them.

When it comes to laypersons’ truth-value judgments regarding modal utterances we investigated in Experiment 1, overall, the data bring very little to no support to our hypotheses, which, again, is good news for two-dimensional semantics. Contrary to H1, we found that the majority of lay speakers tend to judge modal statements such as ‘Although water is H₂O, water could have a different chemical structure, and be XYZ’ false—this was the case in three out of four experimental conditions in our study. Notably, participants were more likely to deny the claim that cats could have been mammals when it was uttered in a world where all cats were robots than to accept it. However, the same was not the case for an analogous scenario depicting a world where the term ‘water’ referred to XYZ (*Water-Dissimilar* condition); here, the truth-value judgments regarding the modal claim that water could have been H₂O split in half. Nevertheless, this is the only

result yielded by our study that poses a problem for the predictions of two-dimensional semantics.

Experiment 1 found no evidence in support of H2. Unlike the hypothesis suggested, there was no systematic difference in truth-value judgments elicited by the scenarios involving different kind terms ('water' and 'cats'). However, we did obtain some support for H3, since the *World* manipulation did affect participants' judgments in the *Water* case (but, contrary to the hypothesis, not in the *Cats* case).

One possible conclusion is that we found an empirical effect that poses an explanatory challenge to two-dimensional semantics: whereas, according to its predictions, semantic judgments should be shaped by the same set of principles across various possible worlds considered actual, the observed differences between *Water-Similar* and *Water-Dissimilar* conditions indicate that in one of these cases another (perhaps philosophically irrelevant?) factor is at play.

However, there is another option worth considering. Since the *Water-Dissimilar* case is clearly an outlier in our dataset (all the other conditions elicited intuitions, by and large, consistent with two-dimensional semantics), and not only concerning truth-value judgments, but also subjects' confidence in their responses, one might try to explain away this result in terms of a bias. So, here is an attempt to provide an error theory of positive *Water-Dissimilar* truth-value judgments: notice that the belief that water is H₂O is part of widespread, common knowledge, and laypersons are frequently exposed to information that water is H₂O. Hence, when considering the question of whether, in a world where water is XYZ, it could have been H₂O, participants' judgments might be affected by the familiarity bias.⁴ It is a well-documented phenomenon that the familiarity bias influences judgments regarding the likelihood of events: some events are considered more likely than others only because the former are more familiar than the latter (e.g., Tversky & Kahneman, 1973). The same could be the case for modal judgments about possibility.

This idea led us to conduct a follow-up study, where we adopted similar methods as in Experiment 1, but employed a scenario that involved a kind whose essential property (also chemical structure, similarly to the *Water* scenario) is not well-known: ethanol.

⁴ On the other hand, the concept of robotic cats should be readily accessible to our participants, and not only due to frequent exposure to representations of artificial pets in science-fiction literature and cinema, but also thanks to *actual* attempts at manufacturing robotic pets, which are widely advertised and publicized.

5. Experiment 2

In the follow-up study, we tested whether replacing the kind WATER in the study materials with another, similar kind, whose chemical structure, unlike in the case of water, is not commonly known, would result in removing the impact of *Similar/Dissimilar* factor on truth-value judgments that we observed in Experiment 1 for the *Water* scenario. For this purpose, we constructed a scenario structurally similar to the *Water* case, where the kind under consideration was ETHANOL (with the chemical structure either C_2H_5OH or T_3Ka_6SY).

5.1. Methods: participants, materials, and procedure

The methods employed in the follow-up study were analogous to those used in Experiment 1, with only one scenario (*Ethanol*) and two possible world conditions: *Similar* and *Dissimilar*. The study was also conducted online. Similarly as in the case of the original experiment, the case description presented to participants started with an introduction mutual to both conditions:

Here are two different kinds of colorless liquids with a pungent taste and intense smell, which make people drunk when consumed, but differ with respect to their chemical structure:

- a) C_2H_5OH [the picture used for the *Water* case in Experiment 1]
- b) T_3Ka_6SY [the same picture as in (a)]

C_2H_5OH and T_3Ka_6SY look, smell, and taste exactly the same and have the same effect when consumed by living creatures.

Then, depending on random assignment to one of two conditions, every participant read the following story in one of two variants:

Ethanol: Similar/Dissimilar

Imagine a world that is pretty much like ours. [What]_{similar} / [However, the history of this world took a different turn, and, as a result, what]_{dissimilar} can be naturally produced by the fermentation process of sugars by yeasts, and can be used as a disinfectant or as an ingredient of drinks used for recreational purposes has the structure [C_2H_5OH]_{similar} / [T_3Ka_6SY]_{dissimilar} and is called alcohol [, just as is the case in our world.],_{similar} / [.]_{dissimilar} All alcohol is exactly like that. There is no fluid of the structure [T_3Ka_6SY]_{similar} / [C_2H_5OH]_{dissimilar} in this world.

Every respondent answered the same set of questions as in Experiment 1. Additionally, at the end of the survey, they were asked: ‘Did you know that the chemical structure of ethanol actually is C_2H_5OH before taking this survey?; [Yes–No].’ This allowed us to test whether the participants had prior knowledge regarding the chemical structure of ethanol. The survey included the same attention check as in Experiment 1 and concluded with the demographic section.

Since our additional data collected in the follow-up were analyzed against the results of Experiment 1 obtained for the *Water* case, the design of the study can again be described as a 2×2 between-subjects design.

We hypothesized that, unlike for the *Water* case, we would not observe a significant impact of world similarity on participants’ truth-value judgments regarding the target modal claim for the *Ethanol* scenario. We predicted to find that the majority of respondents would judge the modal claim false in both *Ethanol-Similar* and *Ethanol-Dissimilar* conditions, similarly to what we observed for the *Cats* scenario in Experiment 1.

Similarly to Experiment 1, the follow-up study was preregistered and the preregistration was published on the OSF webpage (<https://osf.io/2rtua>).

We collected responses from 167 Prolific users, but 46 were excluded from the final analysis due to failed attention or comprehension checks. Together with the data concerning the *Water* scenario from Experiment 1, the number of observations for the analysis was $N = 248$. Participants’ age ranged from 20 to 77 years ($M = 40.72$; $SD = 13.04$). 51.6% of them self-identified as male, 46.4% female, 1.2% non-binary, and 0.8% refused to provide information about gender.

5.2. Experiment 2: Results

5.2.1. Results: truth-value judgments

Before we report the main result of the follow-up study, it is worth mentioning that no participant declared prior knowledge of the molecular formula of ethanol, so our assumption that this piece of information is not common knowledge proved legitimate.

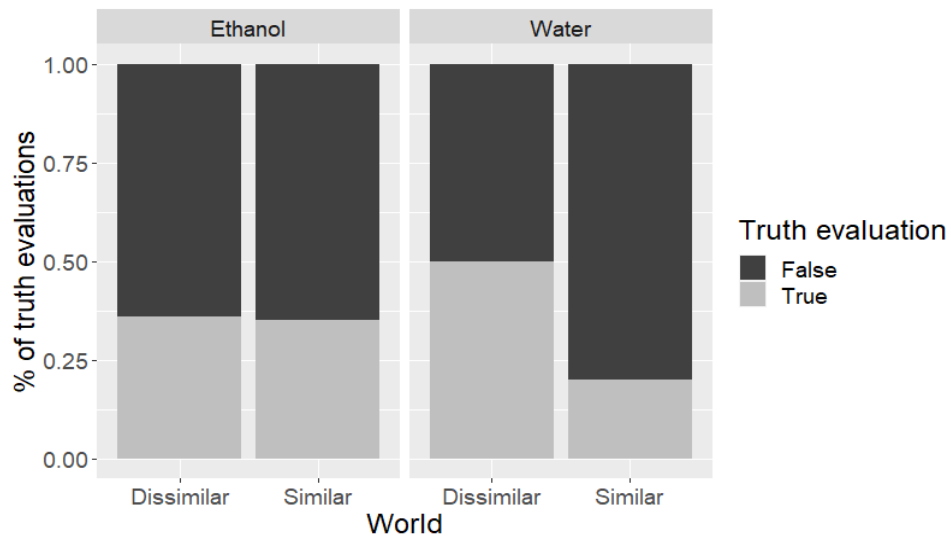
For a general overview of the findings regarding respondents’ truth-value judgments of the target modal claim, we subjected the data to LOGIT analysis (which is summarized in Table 2).

Table 2. Logistic regression for truth-value judgments in Experiment 2: main effects and interaction.

Predictor	<i>B</i>	<i>SE B</i>	<i>z</i> -value	<i>p</i> -value
Scenario				
Ethanol	—	—	—	—
Water	0.57	0.36	1.61	0.107
World				
Similar	—	—	—	—
Dissimilar	-0.05	0.38	-0.12	0.903
Scenario × World				
Ethanol – Similarity	—	—	—	—
Water – Similarity	-1.34	0.56	-2.40	0.017
Constant	-0.57	0.27	-2.15	0.032

Similarly to Experiment 1, we did not observe a main effect of *Scenario* ($p = 0.11$) or *World* manipulation ($p = 0.90$), but a significant interaction between the two factors did emerge ($B = -1.34$, $z = -2.40$, $p = 0.017$). Recall that for the original *Water* case, the *World* manipulation had an impact on truth-value judgments—participants were more likely to judge the modal claim true in the *Dissimilar* condition than in the *Similar* condition (see section 3.3.1). However, the follow-up found no influence of world similarity on truth-value judgments elicited by the *Ethanol* scenario. More specifically, a significant majority of participants (binomial tests, two-sided) judged the target modal utterance in the *Ethanol* case false in both *Similar* ($p = 0.027$) and *Dissimilar* ($p = 0.039$) conditions, which mirrors the results we obtained in Experiment 1 for the *Cats* case. This result is illustrated in Figure 5.

Fig. 5. The proportion of truth-value judgments depending on the experimental condition, *Ethanol* vs. *Water* scenario.

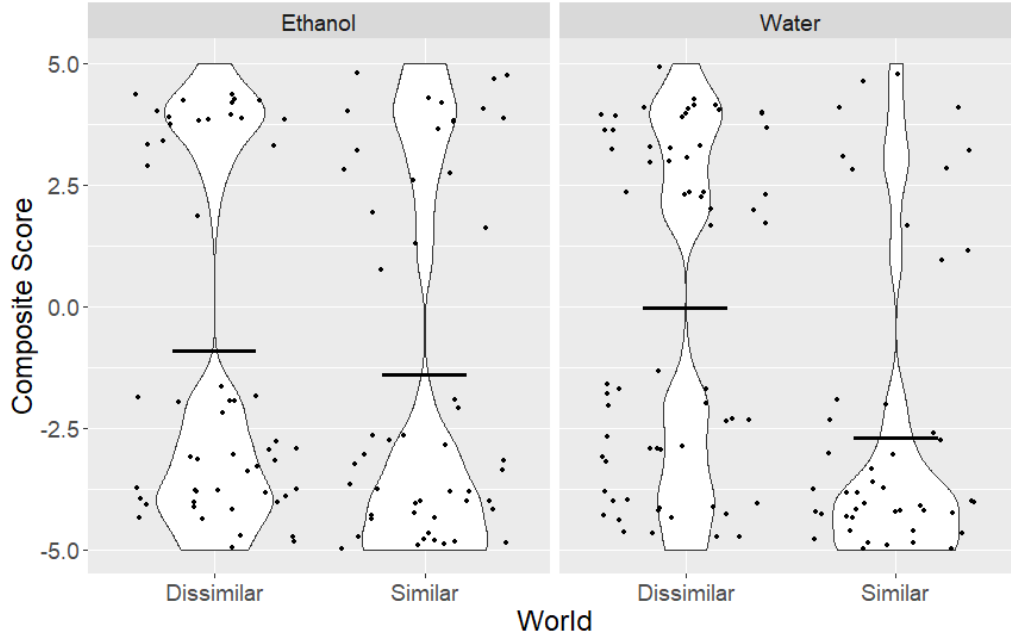


5.2.2. Results: composite scores

The overall results yielded by combining truth-value judgments with confidence ratings paint a more nuanced picture to what was observed for truth-value judgments alone since, although the 2×2 ANOVA analysis found no main effect of *Scenario* ($F(244, 1) = 0.01$; $p = 0.94$, $\eta^2 < 0.001$), it did detect a main effect of *World* manipulation ($F(244, 1) = 12.22$; $p < 0.001$, $\eta^2 = 0.048$) with overall lower scores in the *Similar* condition ($M = -2.03$; $SD = 3.54$) than in the *Dissimilar* condition ($M = -0.44$; $SD = 3.67$), and a significant interaction between the two factors ($F(244, 1) = 5.62$; $p < 0.019$, $\eta^2 = 0.023$).

Further post-hoc pairwise comparisons (Tukey HSD) found no difference in participants' composite scores between the *Similar* ($M = -0.92$; $SD = 3.73$) and *Dissimilar* ($M = -1.42$; $SD = 3.83$) condition for the *Ethanol* scenario ($p = 0.87$), but for the *Water* scenario the average composite score was significantly lower in the *Similar* ($M = -2.71$; $SD = 3.09$) than in the *Dissimilar* ($M = -0.04$; $SD = 3.60$) condition ($p < 0.001$). The results are summarized in Figure 6.

Fig. 6. The distribution of composite scores depending on the experimental condition, *Ethanol* vs. *Water* scenario. Negative values indicate negative verdicts regarding the target modal utterance, positive values indicate the opposite. Vertical bars represent the average score.



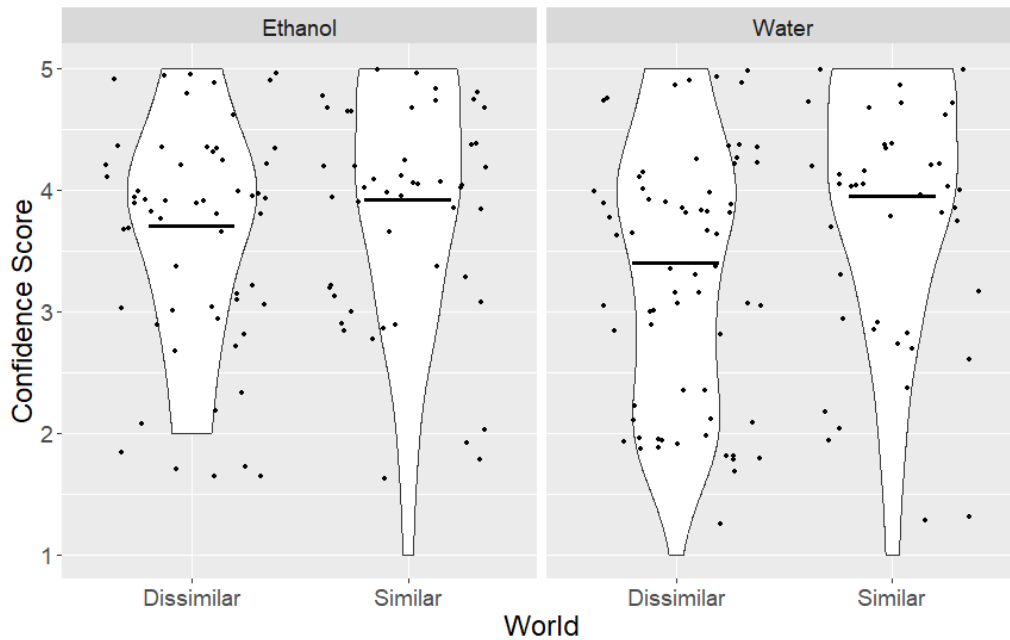
5.2.3. Results: confidence ratings

When it comes to confidence scores, the analysis only found a significant main effect of *World* manipulation ($F(244, 1) = 8.14$; $p = 0.005$, $\eta^2 = 0.032$) but no effect of *Scenario* ($F(244, 1) = 1.69$; $p = 0.195$, $\eta^2 = 0.005$) or interaction ($F(244, 1) = 1.55$; $p = 0.215$, $\eta^2 =$

0.006). Overall, participants felt more confident when evaluating the *Similar* world ($M = 3.93$; $SD = 1.07$) than the *Dissimilar* counterpart ($M = 3.54$; $SD = 1.03$).

Recall that in Experiment 1, we found a significant drop in participants' confidence with their answers in the *Water-Dissimilar* condition compared to the other experimental conditions (see section 3.3.3). Interestingly, an analogous effect did not emerge for the *Ethanol* case, where no significant difference in average confidence ratings was found between the *Similar* ($M = 3.92$; $SD = 1.06$) and *Dissimilar* ($M = 3.70$; $SD = 0.92$) conditions (Tukey HSD, $p = 0.68$). The results are shown in Figure 7.

Fig. 7. The distribution of confidence ratings depending on the experimental condition, *Ethanol* vs. *Water* scenario. '5' — 'completely certain'; '1' — 'completely uncertain.' Vertical bars represent the average score.



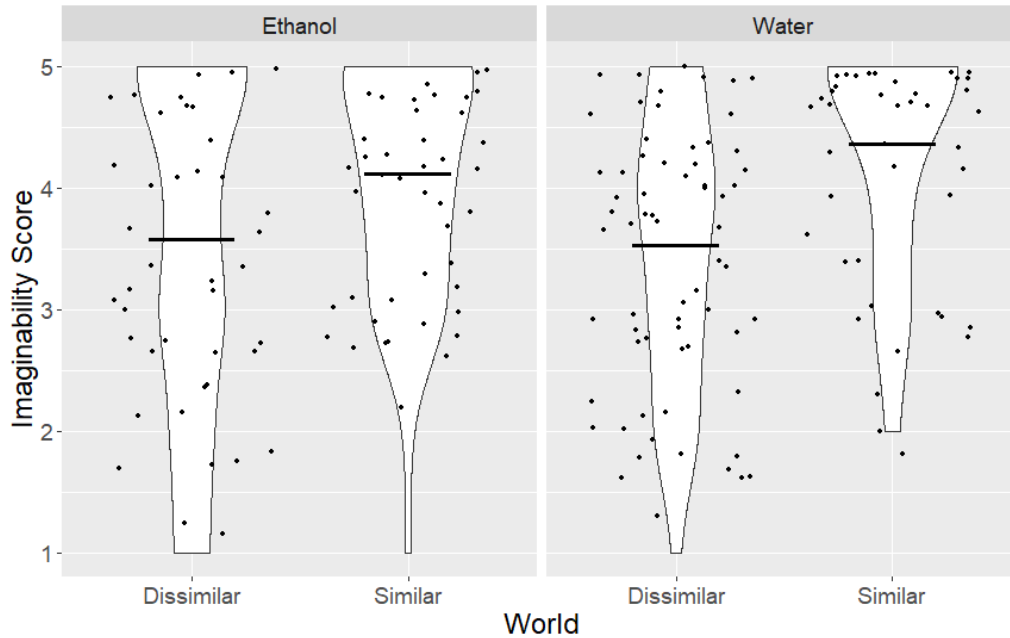
5.2.4. Results: imaginability ratings

With regards to imaginability ratings, similarly as in Experiment 1, in the 2×2 ANOVA analysis for the follow-up study no main effect of *Scenario* ($F(244, 1) = 0.11$; $p = 0.746$, $\eta^2 = 0.002$) or interaction ($F(244, 1) = 1.03$; $p = 0.312$, $\eta^2 = 0.004$) was found, and the only effect that emerged was the main effect of *World* manipulation ($F(244, 1) = 22.85$; $p < 0.001$, $\eta^2 = 0.086$). Overall, participants' imaginability scores were higher in the *Similar* ($M = 4.23$; $SD = 0.96$) than in the *Dissimilar* ($M = 3.55$; $SD = 1.26$) condition.

In this respect, the *Ethanol* scenario elicited similar judgments as the *Water* and *Cats* scenarios. Just like in these two cases, when evaluating the *Ethanol* scenario, respondents found the *Dissimilar* ($M = 3.57$; $SD = 1.42$) world harder to imagine than

its *Similar* ($M = 4.36$; $SD = 0.95$) counterpart (Tukey HSD, $p = 0.044$). See Figure 8 for detailed results.

Fig. 8. The distribution of imaginability ratings depending on the experimental condition, *Ethanol* vs. *Water* scenario. ‘5’ – ‘very easy [to imagine]’; ‘1’ – ‘very difficult [to imagine]’. Vertical bars represent the average score.



6. Theoretical and methodological upshots of the study

Two consequences follow from the study. The first is more general and concerns two-dimensional semantics, its relation to semantic intuitions, and broader issues in the philosophy of language and modal epistemology. The second regards the methodology of experimental philosophy and stems from our follow-up study and the confirmation of the hypothesis concerning the impact of familiarity bias on the results of Experiment 1.

The main upshot of the study for modal semantics is that folk intuitions about modality are not in conflict with the predictions of two-dimensional semantics. Taking into account the results of the follow-up experiment, which allow us to explain away the seemingly problematic data yielded by the *Water-Dissimilar* condition in Experiment 1, none of the evidence we gathered shows that folk intuitions are inconsistent with two-dimensional semantics. In addition, it should be noted that our study not only provided evidence that intuitions are consistent with the predictions of two-dimensional semantics, but also strengthened the case for this semantic theory by showing stability of this pattern of responses across various natural kind terms and different essential properties at different worlds. Thereby, although it does not by itself provide evidence to favour two-dimensional

semantics over one-dimensionalism, it places a strong explanatory burden on any error-theory aimed at explaining away the observed pattern of intuition.

One possible objection one might raise is that, although the data we gathered show that a significant majority of lay speakers have intuitions that are in line with the predictions of two-dimensional semantics and stable across different cases, it should be pointed out that a noticeable portion of our participants expressed intuitions to the contrary. In other words, we did not find an effect where nearly all or even 90% of participants provided the same answers to the scenarios. This might suggest that intuitions are not unanimously consistent with the predictions of two-dimensional semantics.

On the one hand, this does not seem to be a serious problem for our interpretation of the data, given that it is extremely rare to find results in experimental philosophy showing agreement in participants' responses even remotely close to 90%. Moreover, our interpretation does not hinge exclusively on the majority of responses to modal claims but also on the stability of these responses across different scenarios and the observed effect in truth ascriptions in modal judgments concerning the world-manipulation that aimed to represent differences between primary and secondary intensions. Therefore, the study overall provides more nuanced evidence in favor of the claim that the predictions of two-dimensional semantics are reflected in semantic intuitions. On the other hand, however, the fact that some respondents had incompatible intuitions should be taken into account, particularly from a broader methodological perspective. This fact should either be addressed by providing an explanation of why some people might have incorrect modal intuitions through a theory concerning semantic competence and its constraints, or by highlighting potential methodological issues with possible limitations in evoking relevant intuitions for philosophical purposes. That being said, this appears to be a general methodological problem rather than a specific issue with our study.

Besides this substantial philosophical finding, there is an important methodological outcome of our study that we would like to highlight. As mentioned, in the first experiment, although we observed intuitions mostly in line with the predictions of two-dimensional epistemic semantics, lay speakers' truth-value judgments regarding target modal claims significantly differed between the *Similar* and *Dissimilar* worlds when the term 'water' was involved. More specifically, in comparison to other experimental conditions, when considering the possible world where the term 'water' referred to XYZ (*Dissimilar* world) lay speakers were more willing to accept the claim that water could have a different chemical structure (H_2O). Importantly, this effect vanished when we used a different term, 'alcohol,' whose actual chemical structure is not widely known.

Laypersons were equally likely to deny the modal claim that alcohol could have a different chemical structure in both *Similar* and *Dissimilar* worlds.

Whether or not our explanation of this phenomenon referring to the familiarity bias is correct, the methodological lesson is clear. The term ‘water’ should be seen as plausibly problematic when used in experimental studies of modal language that significantly involve intuitions about essences, as it elicits different semantic judgments than other expressions that belong to the same class (i.e., natural kind terms). If our hypothesis that familiarity bias is responsible for the differences in semantic judgments we observed for sentences with the term ‘water’ is correct, then generally terms that refer to kinds whose essential features are strongly associated with them should be avoided in similar studies in experimental semantics. This is a methodological desideratum, which is quite different from the one that Devitt and Porter (2021) put forward when arguing that in experimental studies only terms that are well known to the participants should be used—otherwise respondents might not have access to their meanings.

This outcome is crucial when considering the types of terms frequently used in studies in experimental philosophy of language. Due to the importance of Putnam’s Twin-Earth scenario (Putnam 1975) for the philosophical dispute between externalism and internalism, similar scenarios where ‘water’ is supposed to refer to a substance with either the H₂O or XYZ chemical structure are used very often, especially in studies on the externalism/internalism dispute and, more importantly for our current study, psychological essentialism (see e.g., Genone & Lambrozo 2012; Tobia et al. 2020; Haukioja, et al. 2021). In fact, even in our analysis of the state of the art of experimental semantics provided in Section 2, we mentioned several studies that used such scenarios.

In sum, our findings have three consequences for experimental semantics. First, terms like ‘water’ should be avoided in materials of experimental semantics studies for which essential features are crucial (e.g., in studies on essentialism, externalism, or experiments on modal semantics similar to ours). Second, existing results of studies that employed Twin-Earth-like scenarios, which involved the term ‘water’ and significantly relied on its chemical structure as the purported essential property, should be replicated with materials using different expressions. Third, we call experimental philosophers to be mindful of the possible distortion of the data by the familiarity bias in similar studies when using names with well-known essential properties.

7. Conclusion

The study aims to enrich the discussion on modal semantics and modal epistemology with empirical data. We investigated semantic intuitions and their

coherence with the predictions of two-dimensional epistemic semantics. The study highlighted several key points that contribute to these fields.

Most importantly, our findings offer tentative empirical support for two-dimensional epistemic semantics. We observed that laypersons are capable of conceiving of possible worlds in which a given kind term (e.g., ‘cats’) refers to objects possessing different essential features than the essential features shared by what is picked out by this term in the actual world (e.g., imagining a possible world where all cats are robots). Moreover, we found that laypersons generally judged modal statements that entities which belong to a certain natural kind could alter their essential properties to be false, following what essential properties that natural kind has in a world considered to be actual.

We found an exception for this latter phenomenon in claims involving the term ‘water,’ where we observed a difference in lay speakers’ modal judgments between *Similar* and *Dissimilar* worlds. However, we explained it by the familiarity bias and supported our explanation with the results of our follow-up study, where cases were presented in a way designed to be immune to familiarity bias. When we used the term ‘alcohol,’ which refers to a substance with a less well-known chemical structure than ‘water’ (whose chemical structure is widely known), participants’ modal judgments were uniform across *Similar* and *Dissimilar* conditions. Overall, the results for the *Cat* and *Ethanol* scenarios were highly consistent. This fact strengthens the idea that two-dimensional epistemic semantics provides a reliable framework for interpreting modal claims about essential properties.

A methodological upshot of our study is the problematic nature of the common practice of using the term ‘water’ in experimental semantics, particularly in studies that strongly involve beliefs about essential features of objects. Due to its strong association with having an H_2O chemical structure, ‘water’ tends to elicit different modal judgments compared to other natural kind terms. We suggest that this is due to the impact of familiarity bias on participants’ judgments, a hypothesis supported by the results of the follow-up study. Our findings suggest that to obtain more reliable and unbiased results, the term ‘water’ should be avoided in experimental materials of studies concerning essentialism and semantic externalism. Moreover, our results call for revisiting and potentially rerunning previous studies on semantic intuitions, especially those relying on Twin-Earth-like scenarios that involve this term.

The study highlights the need for continued empirical research on modal intuitions. Although our study represents an initial step into this domain of investigation and only provides tentative evidence in support of two-dimensional semantics, it allows for modest optimism regarding the coherence of folk intuitions with the two-dimensional epistemic framework. We hope that our study will open

a new chapter in experimental philosophy, encouraging further empirical investigations into the nature of modal intuitions.

Acknowledgments

Early drafts of this paper were presented in the Epistemology of Modality seminar at The University of California, Los Angeles, and the 4th European Experimental Philosophy Conference (Jagiellonian University, Kraków). We are grateful to participants of these events for helpful discussions; in particular, we would like to thank Stephen Biggs, Vilius Dranseika, Maciej Tarnowski, Tom Wysocki, Kevin Reuter, and Jakub Krawczyk for valuable comments and suggestions.

Declarations

Preregistration. The preregistration of the two studies is publicly accessible on the OSF website: <https://osf.io/2txpa>; <https://osf.io/2rtua>

Data availability. The data that support the findings of this study are openly available in the OSF repository, <https://doi.org/10.17605/OSF.IO/QW9J6>

Funding. The study was financially supported by the Faculty of Philosophy, University of Warsaw (a research subvention granted to Adrian Ziółkowski) and the National Science Centre, Poland (grant no. 2023/51/B/HS1/02050 awarded to Krzysztof Sękowski).

Author contributions. The order of authors was decided through a mutual agreement and reflects the amount of work each author put into the project. The original idea for the study was conceived by Anand Vaidya in 2008 during the NEH Summer Institute on Experimental Philosophy at the University of Utah. The study design, materials, and hypotheses were prepared by Krzysztof Sękowski and Adrian Ziółkowski and later discussed and revised by all the authors. The online surveys were prepared by Krzysztof Sękowski. Data collection and statistical analysis were conducted by Adrian Ziółkowski. All the authors contributed to drafting and revising the manuscript. The authors agreed to share the credit for the study in the following way: Adrian Ziółkowski 40%; Krzysztof Sękowski 40%; Anand Vaidya 20%.

References

- Barnes, J., and Black, J. 2016. Impossible or Improbable: The Difficulty of Imagining Morally Deviant Worlds. *Imagination, Cognition and Personality* 36(1): 27-40.
- Chalmers, D. 1996. *The Conscious Mind*. New York: Oxford University Press.
- Chalmers, D. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research* 59(2): 473–496.

- Chalmers, D. 2002. Does conceivability entail possibility? In *Conceivability and possibility*, eds. T. Gendler and J. Hawthorne, 145-200. New York: Oxford University Press.
- Chalmers, D. 2004. Epistemic two-dimensional semantics. *Philosophical Studies* 118(1-2): 153–226.
- Devitt, M., and Porter, B. C. 2021. Testing the reference of biological kind terms. *Cognitive Science* 45(5): e12979.
- Garcia-Carpintero, M., and Macias, J. 2004. *The Two-Dimensionalist Framework: Foundation and Applications*. New York: Oxford University Press.
- Genone, J., and Lombrozo, T. 2012. Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology* 25(5): 717-742.
- Haukioja, J., Nyquist, M., and Jylkkä, J. 2021. Reports from Twin Earth: Both deep structure and appearance determine the reference of natural kind terms. *Mind & Language* 36(3): 377-403.
- Icard, T., Kominsky J., and Knobe J. 2017. Normality and actual causal strength. *Cognition* 161: 80-93.
- Jylkkä, J., Railo, H., and Haukioja, J. 2009. Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology* 22(1): 37-60.
- Kaplan, D. 1989. Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and other Indexicals. In *Themes From Kaplan*, eds. Joseph Almog, John Perry, and Howard Wettstein, 481-563. Oxford University Press.
- Knobe, J. 2021. Philosophical intuitions are surprisingly stable across both demographic groups and situations. *Filozofia nauki* 29(2): 11-76.
- Knobe, J. 2023. Difference and robustness in the patterns of philosophical intuition across demographic groups. *Review of Philosophy and Psychology* 14(2): 435-455.
- Knobe, J. (forthcoming). Conflicting Intuitions. *Ergo*.
- Knobe, J., Prasada, S., and Newman, G. E. 2013. Dual character concepts and the normative dimension of conceptual representation. *Cognition*: 127(2): 242-257.
- Knobe, J., and Szabó, Z. G. 2013. Modals with a taste of the deontic. *Semantics and Pragmatics* 6(1): 1-42.
- Kripke, S. 1980. *Naming and Necessity*. Harvard University Press.

- Liao, S., Strohming, N., and Sripada, C. S. 2014. Empirically Investigating Imaginative Resistance. *British Journal of Aesthetics* 54(3): 339-355.
- Machery, E. 2017. *Philosophy within its proper bounds*. Oxford University Press.
- Machery, E., Mallon, R., Nichols, S., and Stich, S. P. 2004. Semantics, cross-cultural style. *Cognition* 92(3): B1-B12.
- Machery, E., and Seppälä, S. 2011. Against hybrid theories of concepts. *Anthropology and Philosophy* 10: 99-126.
- Nichols, S., Pinillos, N. Á., and Mallon, R. 2016. Ambiguous reference. *Mind* 125(497): 145-175.
- Putnam, H. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131-193.
- Quillien T. 2020. When do we think that X caused Y? *Cognition* 205: 104410.
- Salmon, N. 1989. The logic of what might have been. *Philosophical Review* 98(1): 3-34.
- Sękowski, K., Ziółkowski, A., and Tarnowski, M. 2023. Western skeptic vs Indian realist. Cross-cultural differences in Zebra Case intuitions. *Review of Philosophy and Psychology* 14(2): 711-733.
- Soames, S. 2005. *Reference and description*. Princeton University Press.
- Stalnaker, R. 1978. Assertion. In *Syntax and Semantics 9: Pragmatics*, ed. Peter Cole, 315-332. New York Academic Press.
- Stich, S. P., and Machery, E. 2023. Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology* 14(2): 401-434.
- Tobia, K. P., Newman, G. E., and Knobe, J. 2020. Water is and is not H₂O. *Mind & Language* 35(2): 183-208.
- Tversky, A., and Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5(2): 207-232.
- Vaidya, A. 2008. Modal Rationalism and Modal Monism. *Erkenntnis* 68(2): 191-212.
- van Dongen, N., Colombo, M., Romero, F., and Sprenger, J. 2021. Intuitions about the reference of proper names: A meta-analysis. *Review of Philosophy and Psychology* 12(4): 745-774.
- Yablo, S. 1993. Is Conceivability a Guide to Possibility? *Philosophy and Phenomenological Research* 53(1): 1-42.