Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-025-00769-w

Conceptual Revision in Action

Ethan Landes[1] · **Kevin Reuter**[2,3]

Accepted: 3 March 2025

Abstract
Conceptual engineering is the practice of revising concepts to improve how people talk and think. Its ability to improve talk and thought ultimately hinges on the successful dissemination of desired conceptual changes. Unfortunately, the field has been slow to develop methods to directly test what barriers stand in the way of propagation and what methods will most effectively propagate desired conceptual change. In order to test such questions, this paper introduces the masked time-lagged method. The masked time-lagged method tests people's concepts at a later time than the intervention without participant's knowledge, allowing us to measure conceptual revision in action. Using a masked time-lagged design on a content internalist framework, we attempted to revise PLANET and DINOSAUR in online participants to match experts' concepts. We successfully revised PLANET but not DINOSAUR, demonstrating some of the difficulties conceptual engineers face. Nonetheless, this paper provides conceptual engineers, regardless of framework, with the tools to tackle questions related to implementation empirically and head-on.

1 Introduction

Conceptual engineering is the practice of improving concepts so that we can think and talk better.[1] Rather than merely trying to understand existing conceptual or linguistic structures, conceptual engineers aim to be conceptually or linguistically creative, revisionary, and innovative in order to address shortcomings in our current conceptual

[1] For the rest of this section, we use "concept" in the broadest possible sense to include any potential target of conceptual engineering. Starting in Section 2, we use "concept" specifically in the content internalist sense to refer to token cognitive entities.

We would like to thank Pascale Willemsen, Manuel Gustavo Isaac, Eugen Fischer, Lucien Baumgartner, commenters on the XPhi Blog and anonymous referees for comments on earlier versions of this work, and the editor of Review of Philosophy and Psychology for suggesting the follow-up experiment. We would also like to thank the audiences of the Dawn of Experimental Conceptual Engineering workshop at the University of Zurich, ECAP at the University of Vienna, the Joint Session at Birkbeck, the Concepts and Philosophical Methodology workshop at the University of Cambridge and the conceptLab Seminar at Hong Kong University for their questions and comments.

Extended author information available on the last page of the article

Published online: 26 April 2025

schema. Using the four-part process of Isaac et al. (2022), conceptual engineering involves *describing* people's existing concepts, *evaluating* whether they fall short in some way, *improving* upon them by designing a new or revised concept, and *implementing* the product by spreading it to the right people.

What exactly one should take the target of conceptual engineering to be is an open and contentious issue in the conceptual engineering literature (Belleri 2021; Isaac et al. 2022; Sawyer in press). Irrespective of this unresolved question, however, the first three steps of conceptual engineering involve processes that are well-understood by academics. Cognitive science, developmental science, experimental philosophy, armchair philosophy, and other fields have long been in the business of describing people's concepts, such as CAUSE, TRUTH and RESPONSIBILITY, using methods like thought experiments, laboratory demonstrations, corpus linguistics, etc. Similarly, philosophy and other fields have long been in the business of the second step of conceptual engineering, evaluation, when they use thought experiments, arguments, formalizations, and other tools to determine whether concepts lead to fallacies or other unwanted consequences. Improving the concept, the third step of conceptual engineering, has a less storied history, but recent philosophers have drawn lessons from design and engineering and have developed improved concepts themselves (Haslanger 2000; Napolitano and Reuter 2023; Reuter and Brun 2022; Scharp 2013). What about the fourth and final step of the conceptual engineering process?

1.1 Local and Global Implementation Questions

The final step of the process of conceptual revision concerns the implementation of engineered concepts. Here, conceptual engineers face two types of empirical questions:

- GLOBAL IMPLEMENTATION QUESTIONS: What factors *in general* affect the success or failure of propagating concepts?
- LOCAL IMPLEMENTATION QUESTIONS: What factors affect the success or failure of revising or replacing *specific* concepts?

Conceptual engineers have hitherto tried to answer global implementation questions by either drawing parallels from historical examples of conceptual propagation and linguistic change or by drawing lessons from cognitive science. The historical case studies start from a straightforward premise: linguistic change and collective conceptual change have happened in the past, and lessons can be learned from those case studies (Koslow 2022; Landes in press; Thomasson 2021). For example, Landes (in press) examines the problems faced by public health officials when SOCIAL DISTANCING was propagated worldwide in early 2020, arguing the label "social distancing" potentially hindered accurate conceptual propagation.

Others have tried to make ground on global implementation questions by looking towards cognitive science and arguing that cognitive features make some implementations easier than others. Machery (2021) combines experimental philosophy work on

the defectiveness of INNATE (Griffiths and Machery 2008; Machery et al. 2019) with cognitive science work on cultural evolution (Buskell 2017; Scott-Phillips et al. 2018; Sperber 1996). Some concepts, such as INNATE, will not easily be revised, Machery argues, because we are naturally attracted to features of the concept. Fischer (2020) instead focuses on what work on the cognitive structure of polysemy can teach about conceptual revision. Drawing on work showing certain types of polysemy can lead to conflation (Fischer and Engelhardt 2019; Fischer et al. 2015; Giora 2003), Fischer argues that conceptual engineers should avoid novel senses of words that require us to suppress features of more dominant senses.

Local implementation questions have gained less attention. While changes have been proposed by conceptual engineers, very little effort has been made to actually propagate them. Conceptual engineers have come up with various proposed designed changes, such as splitting TRUTH (Scharp 2013), ameliorating WOMAN (Haslanger 2000), and introducing "conspiratorial explanation" as a neutral counterpart to "conspiracy theory" (Napolitano and Reuter 2023). These have, however, merely remained proposals. In contrast, the examples we do have of successfully propagated designed cognitive and linguistic devices — the sex/gender distinction (Muehlenhard and Peterson 2011), the revision of PLANET (IAU 2006), and the relabeling of "gross stress reaction" to "Post-Traumatic Stress Disorder" (Andreasen 2010; Saigh and Bremner 1999) — have all come from outside of the academic discipline of conceptual engineering. At the time of writing, self-identifying conceptual engineers have largely limited their efforts at propagating their designed changes to presenting the changes in academic articles, and they have been slow to research what sorts of propagation efforts are required to get their designed changes to stick.

Ultimately both global and local implementation questions require an empirical approach, regardless of the conceptual engineering framework (Landes 2025). Whatever the target of conceptual engineering is, something needs to be propagated, whether a cognitive disposition, causal historical chain, speaker meaning, etc. However, since these targets differ significantly, one might assume that each framework presents its own set of empirical inquiries. For instance, what factors influence the success of the spread of a linguistic norm may well differ from the factors that impact the uptake of adjusted categorization patterns. Nevertheless, the purpose of this paper is to offer a modular framework that can be expanded to be used regardless of framework.

1.2 In Need of a New Method

How can conceptual engineers directly test specific hypotheses raised by the global and local propagation questions? Developmental psychology may prove as a source of inspiration, particularly in relationship to global implementation questions. Like conceptual engineers, many developmental psychologists are interested in how concepts change, albeit typically during natural development and education as opposed to in response to conceptual engineering (e.g., Carey 2011; Poling and Evans 2004;

Shtulman and Calabi 2013; Spelke and Tsivkin 2001). Here, the methods of developmental psychology have uncovered lessons relevant to conceptual engineering. For one, naive folk concepts do not morph into scientific concepts, instead the folk concepts appear to be suppressed by the novel scientific concepts (Shtulman and Valcarcel 2012). For another, depending on the concept, conceptual change can take months or even years, and such complex conceptual changes may only be present in a fraction of the adult population (Carey 2011).

There are multiple reasons why the methods of developmental psychology are of limited use to conceptual engineers hoping to answer global and local questions about implementation. First, developmental psychologists primarily study how children and young adults undergo conceptual change, focusing on their ability to learn new concepts and adapt to scientific theories. In contrast, conceptual engineers often aim to implement conceptual change in adults who have held specific concepts for long periods of time. Second, existing empirical studies often examine conceptual change in traditional educational settings such as classrooms. However, conceptual engineers may want to target adult populations and change their views through non-formal means of information dissemination. Third, practical considerations favor survey-based designs over classroom-based and other in-person designs. In general, survey-based designs are easier and cheaper to run than in-person designs. At the same time, due to contingent historical facts, empirically-driven analytic philosophers, as well as other researchers in fields that are often more theoretical in nature, typically do not currently have the logistics or expertise in place to run, for example, in-person demonstration-based or game-based studies — a common design to test for conceptual change in infants and children. The shift to adults and survey-based designs creates issues, however, as surveys on adults are notoriously susceptible to noise caused by participants either misreading the intention of the experimenter or relying on subtle pragmatic cues to interpret questions in unwanted ways (Conrad et al. 2014; Cullen 2010; Schwarz 1995).

In short, an experimental method is needed that (a) targets adult populations, (b) employs non-formal means of information dissemination, (c) uses a survey-based design, and (d) minimizes unwanted survey pragmatics.

One of the central goals of the present paper is to propose masked time-lagged designs as a solution to these four desiderata. Masked time-lagged designs can test people's conceptual understanding at a different point in time as the intervention, without participants' knowledge that the second session is related to the first. This allows the study of adult populations on survey platforms (a & c) with ecological validity (b) while giving researchers powerful control over survey pragmatics (d).

Time-lagging is especially helpful for improving the study's ecological validity of non-formal information dissemination and uptake. To test whether any measured change occurs over the timescale that interests conceptual engineers, data should be collected later than the intervention. Conceptual engineers aim to bring about long-term linguistic or conceptual changes, and it is not enough that people learn novel content for a few minutes and then either forget it or never deploy it again. Accordingly, any design testing conceptual propagation will have to take steps to avoid collecting the contents of short-lived ad hoc concepts or information stored in working memory. Testing revision at separate points in time avoids this issue.

In contrast to time-lagging, masking is particularly valuable at reducing the impact of survey pragmatics and participant mind-reading that may get in the way of measuring conceptual change. Two aspects stand out when considering the importance of masking: First, if the test questions are not masked, it would be fairly obvious to participants what answers the study is trying to cause. This will artificially increase measured rates of revision when participants answer in a way they think is helpful. Second, masking is required to allow for meaningful comparisons between the control and test groups. Participants in the test group in non-masked designs will, through pragmatic mechanisms, interpret the content being measured as part of the question under discussion during the "conversation" of the study. This will lower the validity of between-subject comparisons, as participants in the test group will be driven by pragmatic mechanisms to be more likely to answer according to the content being measured than the control group will.

By employing a masked time-lagged method—which we specify in the next section—we aim to demonstrate that both global and local implementation questions can be directly tested. In the empirical parts of this paper, we will adopt a content internalist framework of conceptual engineering (e.g., Fischer 2020; Machery 2017; Pollock 2021) to introduce the masked time-lagged method as well as the hypotheses and results of our study (Section 2). Content internalist frameworks take concepts to be token cognitive representations that possess structure and influence cognitive processes such as categorization and inference (Machery 2009; Margolis and Laurence 2007). We adopt an internalist framework in part due to methodological convenience—it is much easier to draw inferences from experimental data to conceptual content if one takes content to be entirely grounded in mental states. We also adopt an internalist framework due to theoretical convenience—it allows us to build from the theoretical work of Fischer (2020) and Machery (2017), who offer clear guidance on how to test conceptual content for invariantist internalist conceptual engineering frameworks in survey-based designs. To our knowledge, no comparable proposals exist for other conceptual engineering frameworks. Following the presentation of our findings in Section 2, we introduce a subsequent study (Section 3), whose results offer a preliminary explanation for some of the outcomes observed in our Main Study. In Section 4, we then discuss the consequences of our findings on the conceptual revision, responding to objections and generalizing the methods to other conceptual engineering frameworks.

2 Main Study

In this section, we discuss the selection of concepts, introduce our methods and hypotheses, and present the results of the masked time-lagged study. Hypotheses and methodology were pre-registered with the Open Science Framework. The different surveys, including all the stimuli, that were administered to the participants are available on this online repository.

2.1 Selection of Concepts

In order to demonstrate how conceptual uptake can be directly tested, we selected two concepts, DINOSAUR and PLANET. We aimed to revise them to be in line with recent scientific discoveries. According to common folk wisdom, dinosaurs are extinct and Pluto is a planet. However, in the last few decades scientists have generally come to the consensus that (a) birds are an existing form of dinosaur, and so dinosaurs are not extinct and (b) Pluto is not a planet because, unlike planets, it has not cleared its orbit of debris.[2] This means many folk have a concept DINOSAUR that, inconsistent with expert consensus, excludes birds and a concept PLANET that, inconsistent with expert consensus, includes Pluto. Compared to other potential concepts, we chose DINOSAUR and PLANET because they have three key features:

First, they are real cases backed by relevant experts. It would be unethical to attempt to use our position of authority to propagate concepts that are not in the participants' epistemic interests (Kitsik 2023; Shields 2021, 2023). However, the shift in content of DINOSAUR and PLANET have been widely adopted and advocated for by paleontologists and astronomers, respectively (Brusatte 2017; NASA 2023).[3] Therefore, the choice of DINOSAUR and PLANET allows the study of propagation of content that is accompanied by a clear normative standard for why the content should be adopted and is endorsed both by a majority of relevant experts and by the experimenters.

Second, these are scientific cases. Conceptual change will likely require buy-in by participants. Because concept change presumably requires participants expend the cognitive effort to appreciate the stimuli to the extent that is needed to change concepts, participants need to view any stimuli as truthful or legitimate. Accordingly, the concepts DINOSAUR and PLANET allow us to write stimuli in a way that piggybacks on the prestige and social stature of science, by, for example, using real NASA-generated images of the solar system.[4] The scientific nature of the concepts additionally aids in respecting the epistemic autonomy of participants. Rather than relying on sneaky tricks or rhetoric, the topics easily allow for stimuli that are clearly sourced explanations of the scientific consensus that are no different than popular scientific writing—and when possible we directly quoted popular scientific writing from respectable sources (see Section 2.3). In other words, the scientific material means we as experimenters could transparently present participants with reasons to change how they understood the intension and extension of DINOSAUR or PLANET by presenting information in the same format one might find in Scientific American or Popular Science.

Third, these are live cases as people are still learning and/or adjusting to the decision by the IAU that PLANET excludes Pluto and the recent series of discoveries by paleontologists that birds are a type of dinosaur. Therefore, at the time of data collection,

[2] For a reconstruction of the revision of Pluto through epistemic utility theory, see Egré and O'Madagain (2019).

[3] Note that there are nonetheless a few prominent opponents to the current IAU definition of planet (Chang 2022).

[4] This is not to say testing non-scientific conceptual change is not possible, but to the extent that buy-in is required by participants (which is itself an open empirical question), other concepts will require different ways to establish legitimacy.

a large percentage of the participant pool did not yet have concepts with the content we hoped to revise.[5]

2.2 The Masked Time-Lagged Method

As discussed in Section 1.2, conceptual engineers are in need of an experimental method that uses a survey-based design and minimizes unwanted survey pragmatics in order to check for genuine conceptual change. This requires a two-stage process involving time-lagging and masking. In the *first phase* of the study, participants in the test group were presented with one of four different stimuli involving the concepts PLANET or DINOSAUR (details below). These stimuli informed participants about how experts define PLANET, which excludes Pluto, and how they view DINOSAUR, which includes birds. Importantly, participants in the control group did not receive an intervention task, given that we wanted to measure the effect of conceptual propagation against a representative sample without the respective impact of conceptual change.

In the *second phase*, the follow-up task was opened to participants who took part in the intervention task, as well as to a large sample of new participants (the control group). The follow-up for the test group was made available from 2 hours to 72 hours after the intervention task was posted. The mean time between responses was 13.0 hours, and the median was 5.1 hours. Every effort was made to mask the connection, including posting the follow-up on a different Prolific account, making sure the look and feel of the two surveys differed significantly, as well as masking the test question with filler questions.[6] All participants were randomly assigned to three different measures (SFP, Completion, Categorization), the details of which are specified below.

Thus, every participant saw one of 3 test questions (Fig. 1) in the second phase, but only the participants in the test group were recruited for the first phase and shown an intervention. Consequently, our design was 2 (Concept) x 3 (Intervention) x 3 (Measure) and fully between-subject.[7]

2.3 Intervention Task (First Phase): Stimuli

In order to try to change concepts while also offering initial answers to local implementation questions about DINOSAUR and PLANET, two sets of stimuli were written for each

[5] By our best estimate from the control conditions of the Main Study (see Figs. 2 and 3), at the time of data collection in December 2022, somewhere between 60% to 80% of US and UK Prolific users had the folk concept of PLANET and somewhere between 70% to 100% had the folk concept of DINOSAUR as their default content. This estimate was reached by using the Categorization measure as one bound and the higher of either the Completion or SFP measure as the other bound. The higher of the Completion and SFP measure was used because we believe that these, if anything, underestimate the rate of content among the population (see Section 2.4). In the exploratory follow-up run in November 2024 (see Section 3), 16% of participants endorsed the expert sense of DINOSAUR and 49% endorsed the expert sense of PLANET.

[6] Because of the differences in intervention length and the need to compensate Prolific workers fairly, participants were recruited in 5 different surveys (4 x Test, 1 x Control) posted on Prolific on three consecutive weekdays at either 15:00 or 16:00 UTC. Order was determined by a coin flip, and participants were excluded by Prolific screening from taking more than one survey.

[7] *Measure* is considered an independent variable since we manipulated it directly and assessed its impact.

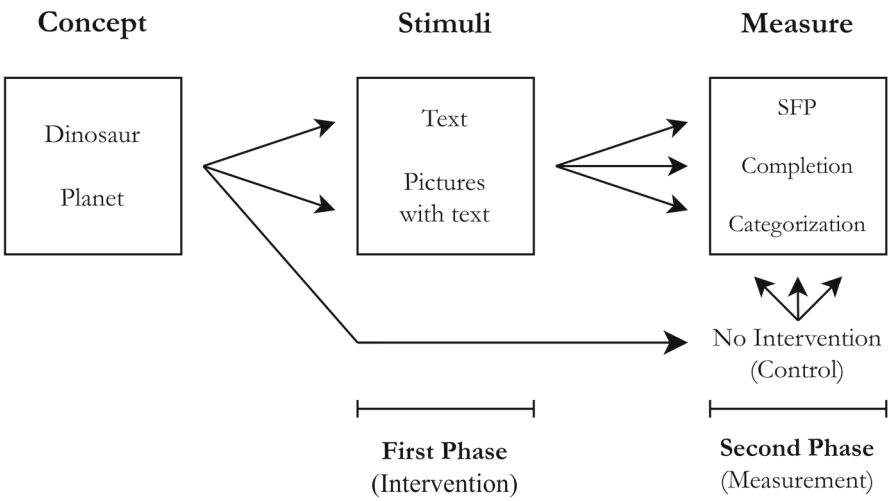

Fig. 1 Diagram of the survey's 2 x 3 x 3 fully between-subject design. Participants were asked questions related to the intervention at t_1, but responses were not analyzed as part of this study. The measure step (t_2) was masked and completed between 1.5 and 73.3 hours after the initial intervention

concept in the implementation task.[8] The first set of stimuli were minimal, text-only interventions (hereafter the "text" interventions). The text interventions were written to explain the conceptual changes behind either DINOSAUR or PLANET in around 200 words and to be readable by participants in one to two minutes. The DINOSAUR text intervention quoted passages from a Scientific American article about the discovery that birds are dinosaurs whereas the PLANET text intervention quoted NASA's version of the IAU definition of planet and quoted a BBC explanation for why Pluto is not a planet. Here is a snippet from the text intervention for DINOSAUR (the full texts can be accessed in the online repository):[9]

> The feathered dinosaurs of Liaoning clinched it: birds really did evolve from dinosaurs. But that statement is perhaps a little misleading because it suggests that the two groups are totally different things. In truth, birds are dinosaurs—they are one of the many subgroups that can trace their heritage back to the common ancestor of dinosaurs and therefore every bit as dinosaurian as Triceratops or Brontosaurus. You can think of it this way: birds are dinosaurs in the same way that bats are a type of mammal that can fly.

The second set of interventions were meant to be more in-depth, multimodal and take 2 to 3 minutes to read (hereafter the "picture" interventions). They used scientific illustrations and photographs alongside text to explain conceptual change and also included quotes from prominent figures stating that birds are dinosaurs or that Pluto

[8] For masking purposes, the interventions included fake test questions that are not included in the analysis below.

[9] Quoted text adapted from https://www.scientificamerican.com/article/how-birds-evolved-from-dinosaurs/

is not a planet. The longer DINOSAUR stimulus's argumentative structure was to try to convince participants that birds and dinosaurs really are of the same kind. The pictures included recreations of prehistoric Theropods, which are notably bird-like, as well as pictures of emus and shoebills, particularly prehistoric looking birds. The longer PLANET stimulus argued that there are two distinct kinds of things that are round and orbiting the sun, and Pluto is more like the non-planets than planets. Images compared the sizes and shapes of Pluto, the Earth, similarly sized objects beyond Neptune, and other celestial bodies, as well as the comparatively messy orbit of Pluto compared to the eight actual planets.[10]

2.4 Measurement Task (Second Phase):

To assess conceptual change, participants from both the test and control groups were presented with the measurement task during the second phase. In order to explore the strengths and weaknesses of different ways of measuring conceptual change, we deployed three measures: COMPLETION tasks, SEMANTIC FEATURE PRODUCTION (SFP) tasks, and CATEGORIZATION tasks. Specifically, the measures were as follows:

Completion: Participants were asked to complete the sentence: "Dinosaurs are ..." or "Pluto is ...", such that a true sentence would be stated.

Semantic Feature Production: Participants were asked to state three characteristics that came to mind when they thought about dinosaurs or Pluto.

Categorization: Participants were asked to categorize whether four items were dinosaurs or planets (the test item being a seagull or Pluto).

Completion tasks are suggested by Fischer (2020) as a way to test for conceptual content. Such conceptual content, at least according to invariantist accounts (compare to contextualists like Casasanto and Lupyan 2015; Yee and Thompson-Schill 2016), is stable over time, stored in long-term memory, and retrieved quickly outside of context (Machery 2009). Fischer (2020) accordingly argues that single-word priming tasks such as "x is ___" are well-suited to get at such content due to their low context and open-ended nature.

Semantic Feature Production (SFP) tasks, sometimes called feature listing tasks, are designed to capture the salient characteristics that participants associate with a given entity. Early examples of such tasks can be found in works by Hampton (1979) and Barsalou (1983), while a comprehensive discussion can be found in Machery (2009) and Reuter (2024). McRae et al. (2005) conducted a study wherein participants were requested to provide salient features associated with hundreds of concepts. Salient features are those that stand out in our mental representation of a particular category compared to other properties. For instance, the feature "dangerous" is considered salient in the concept SHARK, despite the fact that sharks are not necessarily or typically dangerous creatures. Salient features of concepts might be particularly hard to change

[10] Please note that in one of the sentences of the stimuli, we accidentally talk about Pluto's planet instead of Pluto itself.

in people's representation of kinds. Consequently, Semantic Feature Production tasks offer a way to measure how salient features change over time.

Categorization tasks are multiple-choice questions asking participants to select which of several options are members of a given category or kind. Unlike the other two measures, the Categorization task is less likely to under-count the conceptual content of interest because every participant must give a response that bears on the content being studied. Moreover, the Categorization task directly targets what is arguably one of the main functions of concepts—sorting objects into categories. One of the main reasons concepts (again, understood as token cognitive entities) are of interest to conceptual engineers is because they determine how we sort things in the world (Isaac 2023; Machery 2017; Margolis and Laurence 2007), which can in turn influence inferences and decision-making. The Categorization task in particular, however, risks introducing unwanted context by providing implicit contrast cases in the other options.

For all three measures, fillers were implemented as additional strategy to increase the likelihood of successful masking. In the Categorization task, the fillers were other natural kind concepts (e.g. trees), and in the Completion and SFP tasks they were other proper nouns (e.g. Taylor Swift) for PLANET to mask "Pluto" and other kind terms (e.g., chairs) for DINOSAUR to mask "dinosaur". The test question was hidden as one question among five and appeared fourth in a set of five questions. While we did not explain why we were asking the questions we were, when asked, many participants in both the control group and the intervention group guessed the survey purpose involved testing common knowledge or common associations. The exact questions participants saw can be accessed in the online repository.

2.5 Participants & Hypotheses

The masked-time lagged study used a fully between-subject design. 1091 total participants were recruited via Prolific. 361 participants were recruited for the control group, and 730 participants were recruited to the test group and given one of four interventions. Of the 730 participants in the experimental conditions, 560 participants (77%) completed the measurement task, receiving one of the three measures. Of these, 12 participants were excluded for correctly guessing the survey purpose at the end of the post-test, for a final total of 548 (75%). In the control condition, 361 other participants completed one of the six (2 Concepts x 3 Measures) measurement task conditions without any previous intervention. Of the entire sample, the average age was 37.7 (median: 35, min: 18, max: 78), 81% resided in the United Kingdom, 19% resided in the United States, 66% were female, 33% were male, and $< 1\%$ responded other or preferred not to say. The subset of participants in the test group who successfully completed both the intervention task and masked measurement task had a similar demographic composition. The average age was 38.2 (median: 35, min: 18, max: 78), 86% resided in the United Kingdom, 14% resided in the United States, 67% were female, 32% were male, and $< 1\%$ responded other or preferred not to say.

Conceptual engineers disagree about about how easy it is to implement changes (e.g., Cappelen 2018; Jorem 2021; Koch 2021; Nimtz 2024). Therefore, for each of the three measures, our null hypothesis is that there would be no significant difference

between control and test groups, both aggregating the two DINOSAUR and PLANET interventions together and for each of the four interventions separately. For instance, for the Semantic Feature Hypothesis (H1), our pre-registered hypothesis read: There is no significant difference between the test groups and the control group on how often responses referring to being extinct (in the Dinosaur case) or responses referring to planethood (in the Pluto case) are noted in the SFP task.

2.6 Results

The experimenters coded results based on whether participants answered inconsistently with the expert conceptual content. Thus, any participant who wrote "extinct" (or a synonym) for dinosaur or "planet" (or a synonym) for Pluto were counted. Analysis was between-subject, and test scores for each measure were analyzed against the control group.

In the two open-ended tasks, most responses explicitly or implicitly took a stand on dinosaurs' extinction or Pluto's planethood. For dinosaur this included responses like "extinct", "big and extinct", and "birds". For Pluto, results taking a stand on its planethood included responses like "a planet", "a planet in our solar system", or "not a planet". Many responses, especially for DINOSAUR, were orthogonal to the research question, which was an anticipated problem with the open-ended nature of the questions and the reason responses were coded as expert-inconsistent as opposed to expert-consistent. One participant, for example, completed the dinosaur SFP task with "ugly animals", "scaley", "large creatures". While this clearly is eliciting imagery associated with non-avian dinosaurs, the participant does not explicitly take a stand on whether birds are dinosaurs or dinosaurs are extinct, and the participant was therefore coded as not being expert-inconsistent. For Pluto, similarly orthogonal responses included properties like "far" and "cold", as well as participants in both the test and control conditions who interpreted the open-ended questions as being about the Disney character, e.g., "mickey" and "a disny [sic] dog". Because these do not reveal anything about their content of PLANET, these responses were coded as not being expert-inconsistent—unless, as was the case for some participants in the SFP task, they also explicitly endorsed Pluto's planethood in their other listed features. Raw data, coding, and further specifics can be found at the online repository.

Once coded, the two concepts behaved very differently. Starting with DINOSAUR (Fig. 2), combining both the text and picture interventions, there neither was a significant difference between the test and the control group in the Completion task ($\chi^2 = 2.88$, $p = 0.089$) nor in the SFP task ($\chi^2 = 0.77$, $p = 0.38$) (Fig. 2). In the Categorization task, we did, however, see a significant difference between the test and control group ($\chi^2 = 36.61$, $p < 0.001$).

There was no difference between stimuli for DINOSAUR. Among those who saw the text intervention, the Categorization task ($\chi^2 = 35.63$, $p < 0.001$) was statistically significant, but the SFP ($\chi^2 = 1.27$, $p = 0.26$) and Completion tasks ($\chi^2 = 2.59$, $p = 0.11$) were not. Similarly, among those who saw the picture intervention, the Categorization task ($\chi^2 = 28.89$, $p < 0.001$) was statistically significant, but the SFP ($\chi^2 = 0.11$, $p = 0.74$) and Completion tasks ($\chi^2 = 1.70$, $p = 0.19$) were not. Exploratory analysis

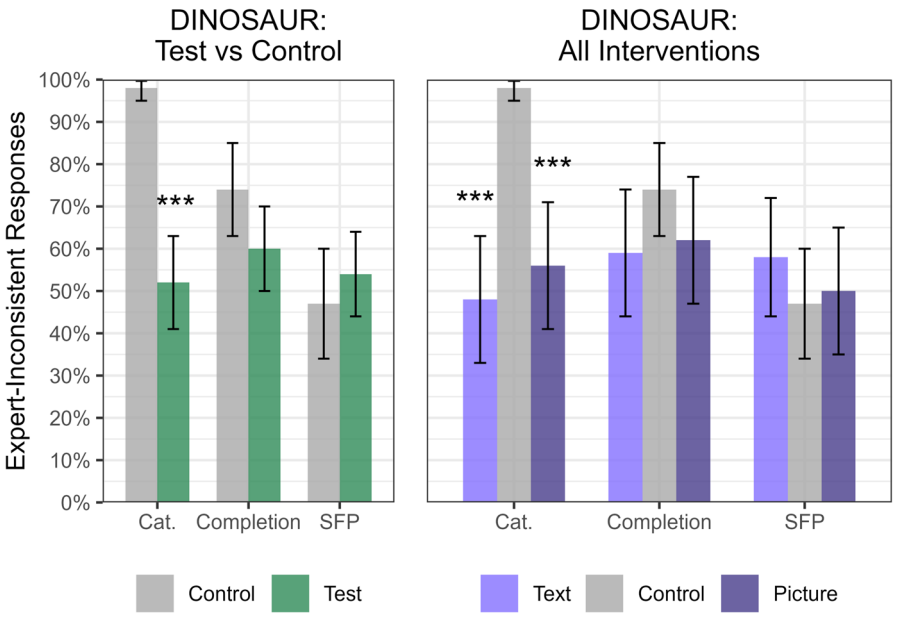

Fig. 2 Results for DINOSAUR grouped as well as broken down by stimuli. The height of the bars show the number of expert-inconsistent responses for the three measurement tasks. Error bars represent 95% confidence intervals

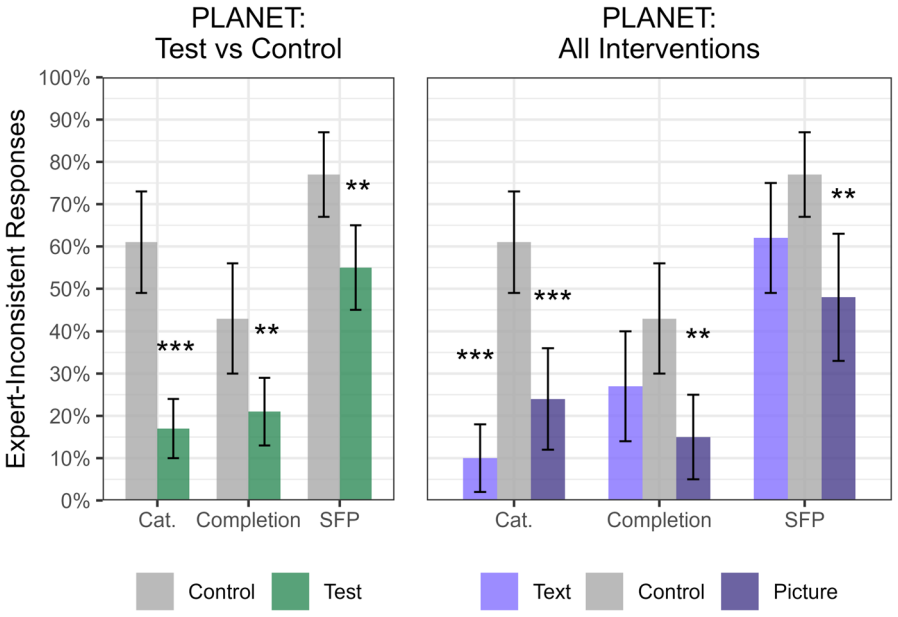

Fig. 3 Results for PLANET grouped as well as broken down by stimuli. The height of the bars show the number of expert-inconsistent responses for the three measurement tasks. Error bars represent 95% confidence intervals

comparing the two stimuli found no significant difference between interventions in any of the three measures (Completion: $\chi^2 = 0.07$, $p = 0.79$; SFP: $\chi^2 = 0.53$, $p = 0.47$; Categorization: $\chi^2 = 0.44$, $p = 0.51$).

In PLANET, results were a bit more promising for conceptual engineering (see Fig. 3). Combining both interventions, the test group was far less likely to say Pluto was a planet than the control group in the Completion task ($\chi^2 = 8.67$, $p = 0.003$), the SFP task ($\chi^2 = 8.05$, $p = 0.005$), and the Categorization task ($\chi^2 = 33.48$, $p < 0.001$). This means that all three null-hypotheses can be rejected as stated in the preregistration.

Collectively, the two sets of PLANET stimuli lowered expert-inconsistent responses on all three measures, suggesting conceptual revision occurred. However, this effect was uneven between the two interventions. The responses of participants who saw the picture stimulus were all statistically significant (Completion: $\chi^2 = 10.14$, $p = 0.001$; SFP: $\chi^2 = 9.98$, $p = 0.002$; Categorization: $\chi^2 = 14.97$, $p < 0.001$), whereas among participants who saw the text stimulus, only the Categorization task was significant (Completion: $\chi^2 = 2.93$, $p = 0.087$; SFP: $\chi^2 = 3.41$, $p = 0.064$; Categorization: $\chi^2 = 31.47$, $p < 0.001$). Exploratory analysis comparing the two stimuli found no significant difference between interventions for any of the three measures (Completion: $\chi^2 = 2.27$, $p = 0.13$; SFP: $\chi^2 = 1.18$, $p = 0.18$; Categorization: $\chi^2 = 3.82$, $p = 0.051$).

2.7 Discussion

Using a masked time-lagged design, our empirical study investigated the possibility of the propagation of conceptual change. By writing stimuli for two distinct concepts, DINOSAUR and PLANET, in two different formats, a short text-only intervention and a longer intervention also containing pictures, we measured the extent to which people's concepts align with the latest scientific discoveries. Our findings suggest that our efforts yielded only limited success in altering people's concept of DINOSAUR. Conversely, we achieved a higher degree of success in revising people's concept of PLANET.

While the Categorization task was significantly different between control and test groups among every intervention for both concepts, this was not the case for the Completion and Semantic feature production tasks. In those measures, after our intervention, people were far less likely to say Pluto was a planet across all three measures compared to our control, but they were just as likely to say that dinosaurs are extinct. It is worth noting that while the length and format of the stimuli appeared to influence participants' responses with respect to PLANET, no comparable effect was discerned for the concept of DINOSAUR.

3 Follow-up Study

The Main Study appears to have successfully revised many participant's conceptual content of PLANET, but the interventions were less effective at altering their conceptual content of DINOSAUR. Naturally, this requires some explanation, and the two concepts are different in a few salient ways. One particularly interesting difference is that PLANET's intension changed as part of an explicit redefinition, while DINOSAUR's

intension remains unchanged—roughly, the creatures that fall under the DINOSAUR evolutionary clade. Examining the impact of intensional and extensional change would require addressing numerous concepts, which we leave to future work. Other plausible explanations include:

(i) Pluto is pretty similar to a typical planet, as it shares several key properties with them—for instance, being a round celestial body that orbits a star. In contrast, birds, and in particular, specific bird species like eagles, are apparently quite dissimilar to typical dinosaurs. These "typical" dinosaurs, such as Tyrannosaurus rex, Brachiosaurus, and Triceratops, are well-known, frequently depicted in media, and representative of major dinosaur groups. It is not implausible to assume that participants are more willing to accept conceptual change when the objects they refer to can more easily be related to their typical instances.[11]

(ii) Changes in DINOSAUR involve expanding the concept to include birds, while changes in PLANET involve shrinking the concept to exclude Pluto (see Liao and Hansen 2023). From very early days in development, we are taught to refine and precisify our concepts. For example, when children first encounter the word "brother", they often overgeneralize it, applying it to any boy they interact with, and only later take into account family relationships (see Ambridge et al. 2013; Rescorla 1980). Consequently, people might find it easier to shrink than to expand concepts.

(iii) The change in PLANET is, at the time of data collection, far more in the cultural zeitgeist than the knowledge that birds are dinosaurs. Many participants may have already partially processed that Pluto is no longer considered a planet but have not had the time or information to partially process that birds are dinosaurs. The explicit teaching we provided on planets may have therefore resonated more strongly with participants than our explicit teaching on dinosaurs.

This follow-up experiment aims to explore the plausibility of the above Conjectures (i), (ii), and (iii). Specifically, the follow-up tests the perceived similarity between Pluto and planets as well as eagles (representing birds) and dinosaurs (Conjecture (i)) and allows us to compare the two (Conjecture (ii)). The study also measures the extent to which participants already know of the technical use (Conjecture (iii)), and whether knowledge of the technical use is tied to perceived similarity. The extent to which there is a difference in similarity and knowledge of the uptake can therefore help inform the extent to which these factors may have influenced the results of the Main Study.

3.1 Participants, Methods & Hypotheses

Through Prolific, 180 participants were recruited with an average age of 38.9 (median: 36, min: 18, max: 72). Of these participants, 56 (31%) identified as men, 119 (66%) identified as women, 3 (2%) identifying as non-binary, and 2 (1%) responding with

[11] A related but distinct hypothesis that will unfortunately have to be tested in future work is that participants see the change in DINOSAUR as violating the epistemic value of homogeneity without gaining an adequate payoff of increased informativeness (see Egré and O'Madagain 2019).

"other" with 141 (78%) residing in the United Kingdom and 39 (22%) residing in the United States. Participants were assigned to either the DINOSAUR or PLANET condition in a between-subject design. Participants first answered three similarity questions in a random order, two of which were fillers. The key similarity questions for participants were as follows:

Dinosaur "How similar is an eagle to a typical dinosaur?"

Planet "How similar is Pluto to a typical planet?"

The filler questions were "How similar is a fig to a typical fruit?" and "How similar is human memory to a typical computer data storage unit?" Participants answered the three questions on a sliding scale between "-3 - Not at all similar" and "3 - Highly similar". Then participants answered a second test question asking whether they agreed with either "Pluto is technically not a planet" or "Eagles are technically dinosaurs", with the options "Yes", "No", and "Don't Know".

With respect to the prediction that Pluto is more similar to the typical planet than birds are to the typical dinosaur, the following hypotheses were preregistered on the Open Science Framework:

Hypothesis 1: There is a significant difference between similarity ratings for eagle/dinosaurs compared to Pluto/planet, such that participants consider eagles to be less similar to the typical dinosaur than they consider Pluto similar to the typical planet.

Hypothesis 2a: Similarity ratings for eagle/dinosaur are significantly below the scale midpoint.

Hypothesis 2b: Similarity ratings for Pluto/planet are significantly above the scale midpoint.

Hypotheses 1 and 2a/b aim to provide support to the idea that similarity to typical instances of a concept might have influenced our results (Conjecture (i)). However, because we predicted prior partial uptake of the revision to impact results (Conjecture (iii)), the following hypotheses were also preregistered):

Hypothesis 3a: Similarity Ratings for eagle/dinosaur are significantly higher for those participants who agree that eagles are technically dinosaurs.

Hypothesis 3b: Similarity Ratings for Pluto/planet are significantly lower for those participants who agree that Pluto is technically not a planet.

3.2 Results

In an unpaired, two tailed t-test, participants rated Pluto to be more similar to the typical planet than eagles to the typical dinosaur, $t(175) = 6.64$, $p < 0.001$, with a large effect size ($d = 1.0$) (Fig. 4). In one sample, one tailed t-tests, the mean rating for eagle's similarity to dinosaurs ($M = -1.17$) was significantly below the scale midpoint of 0, $t(88) = 7.13$, $p < 0.001$, and Pluto's similarity to the typical planet ($M = 0.45$) was significantly above the scale midpoint, $t(87) = 2.51$, $p = 0.007$. The effect size for

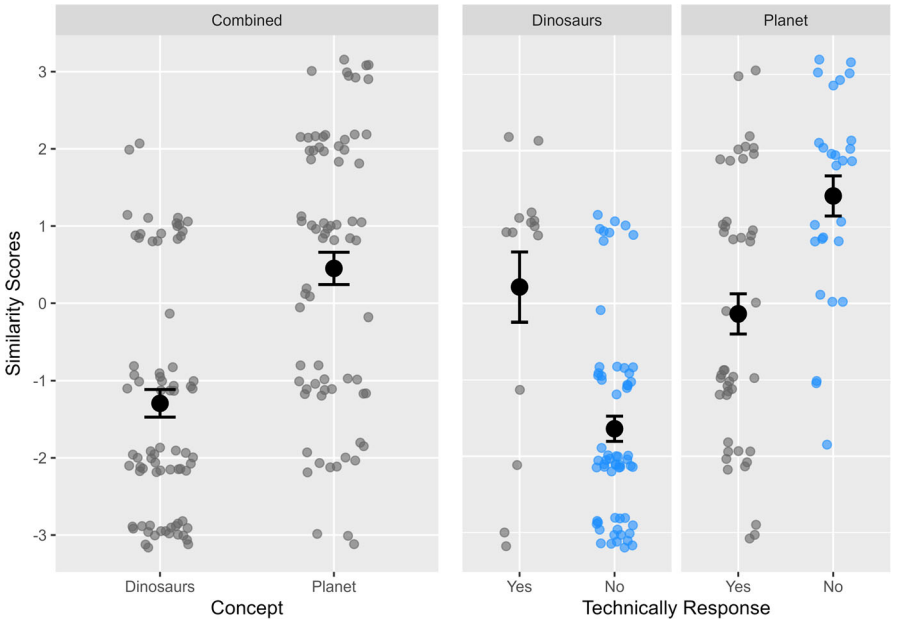

Fig. 4 Results for the similarity scores in the Follow-up Study both combined and broken down by concept and scores in the technically questions. Error bars represent one standard error of the mean

eagle's similarity was medium (d = 0.76) and the effect size for Pluto's similarity was small (d = 0.27). Thus, Hypotheses 1 and 2a/b are supported by our study.

To explore the role of the difference in the structure of conceptual revision—*excluding* Pluto from the set of planets and *including* birds within dinosaurs—the distances of similarity ratings from the scale midpoint for the two concepts were compared as exploratory analysis. A two-sample t-test of similarity responses' mean distances from the scale midpoint showed no significant difference between the conditions, t(174.87) = 1.68, p = 0.095.

Participants were more likely to answer "Yes" than other choices in line with expert usage to the question of whether eagles are technically dinosaurs than if Pluto is technically not a planet ($\chi^2 = 21.39$, $p < 0.001$). In the planet condition, 44 out of 90 (49%) answered "Yes" in line with expert usage when asked whether Pluto was technically not a planet, 28 out of 90 (31%) answered "No", and 18 (20%) answered "Don't Know". In the dinosaur condition, only 14 out of 90 (16%) answered "Yes" in line with expert usage when asked whether eagles are technically dinosaurs, 62 out of 90 (69%) answered "No", and 14 (16%) answered "Don't Know".

To analyze whether agreeing that eagles are technically dinosaurs or Pluto is technically not a planet, the "Don't know" responses were removed from the dataset, and two-sampled one-tailed t-tests compared the mean similarity scores between those that answered "Yes" vs "No". For those in the planet condition, participants that answered "Yes" to whether Pluto was technically not a planet rated Pluto as less similar to the typical planet (M = -0.17) than those who answered "No" (M = 1.41), t(69) = 3.94,

$p < 0.001$ with a large effect size ($d = 0.96$). For those in the dinosaur condition, participants that responded "Yes" to whether eagles are technically dinosaurs rated eagles as more similar to the typical dinosaur ($M = 0.21$) than those that responded "No" ($M = -1.64$), $t(73) = 4.54$, $p < 0.001$ with a large effect size ($d = 1.35$). These results provide support for Hypothesis 3a and Hypothesis 3b.

3.3 Discussion

The interventions of the Main Study were far more successful at revising PLANET to exclude Pluto than revising DINOSAUR to include birds. To explore one particularly plausible reason (Conjecture (i)) why this asymmetry was observed, the Follow-up Study examined the differential similarity between typical planets and Pluto on one side and typical dinosaurs and eagles on the other. The Follow-up Study found that participants see Pluto as being much more similar to a typical planet than eagles are to a typical dinosaur, highlighting one important factor that might have driven the different results we found in the Main Study.

Despite these results, it is important to note that the two cases differ significantly in regards to what they require participants to perform during the implementation process. Specifically, Pluto's perceived similarity to a typical planet may make it harder to *exclude* it from PLANET, just as eagles' lower similarity to a typical dinosaur may make it harder to include them in DINOSAUR. Our analysis showed that the deviation from the scale midpoint was not significantly different between the two cases, suggesting they are on par when considering the asymmetry between exclusion and inclusion processes, leaving us with no clear results in regards to Conjecture (ii).

The Follow-up Study also supports the idea (Conjecture (iii)) that previous exposure to the conceptual change happening to PLANET could be a relevant factor in determining the positive results for PLANET in the Main Study. Not only have more people heard about the planet case before, our results show that previous knowledge changed the extent to which participants rate Pluto similar to a typical planet, as well as rate eagles similar to typical dinosaurs. Since the Follow-up Study was only exploratory, the results should only be seen as a preliminary step toward a more comprehensive explanation of the findings from the Main Study and future research like it.

4 General Discussion

Philosophers have started to think about challenges and obstacles to implementing revised concepts and content (Nimtz 2024; Pinder 2017; Sterken 2020). However, there is a growing recognition that the implementation challenge should be tackled empirically (Andow 2020; Koslow 2022; Landes 2025; Pinder 2017; Thomasson 2021; Wakil 2023), but so far, the field has been slow to develop and deploy the necessary methods. Moreover, those who have used empirical data to study conceptual implementation have not *directly tested* the way specific theories play out in the process of propagating concepts, meanings, or words (Fischer 2020; Koslow 2022; Landes in press; Machery 2021).

To address the current lack of suitable methods for measuring conceptual revision in adult populations, we argued for the use of masked time-lagged designs to directly research the implementation of conceptual change. By testing participants' understanding at two different time points—without revealing that the second stage is connected to the first one—the method increases ecological validity and control over survey pragmatics, making it well-suited for studies on revision conducted on survey platforms. Our findings indicate that the masked time-lagged design offers a credible framework for conceptual engineers seeking to evaluate whether changes to established concepts can be effectively implemented.

Because masked time-lagged designs allow conceptual engineers to directly test questions related to propagation and implementation, the above findings about DINOSAUR and PLANET offer direct, albeit initial, insights into the factors that affect intentional conceptual revision. In Section 4.1, we address several limitations in our methodology before discussing in Section 4.2 how our findings offer first answers to the global and local implementation questions. In Section 4.3, we address the objection that our study measures changes in beliefs as opposed to changes in concepts. In the final section, Section 4.4, we expand the discussion to other conceptual engineering frameworks, arguing that masked time-lagged designs have utility beyond content internalist conceptual engineering frameworks.

4.1 Limitations of Main Study

Before discussing specific inferences that can be drawn about conceptual revision, it is worth acknowledging a few limitations of the Main Study that prevent broader conclusions. Acknowledging these will help map the limits of what inferences can be drawn while also highlighting where future research would be most fruitful.

The first set of limitations is related to the project's scope. Most straightforwardly, only two concepts were investigated. DINOSAUR and PLANET were the only concepts the authors collectively agreed met the ethical and pragmatic desiderata laid out in Section 2.1 and whose change was structured in a way that allowed for measurement using the three measures. Importantly, both PLANET and DINOSAUR are natural kind concepts[12] as opposed to artificial kind concepts like TABLE (see Gelman 2013; Rose and Nichols 2019), abstract concepts like TRUTH, and social kind concepts like JANITOR—not to mention variations of each such as dual character concepts like ARTIST (Knobe et al. 2013; Reuter 2019) and evaluative concepts like VANDAL (Eklund 2011; Willemsen and Reuter 2021). Another limitation of scope is the timeline of measurement. True revisionary uptake would last on the scale of years, not just hours and days, and it is an open question how much uptake of the sort measured above drops off over time. We found no evidence of an effect of time in the dataset, but future exploratory work should explore variations in duration to better understand the effect of time on conceptual uptake based on direct interventions.

[12] These concepts refer to classifications that mirror the inherent structure of the natural world, rather than being driven by the preferences of human beings (Bird and Tobin 2023; Quine 1969).The metaphysics of natural kinds is reflected in our own cognition, as people represent natural kinds as having a core essence responsible for observed superficial similarities (Gelman and Markman 1987; Kornblith 1997).

The second set of limitations is related to the use of a two session between-subject design. This was chosen as the most likely method to result in successful masking. As discussed in Section 1.2, masking allows for valid comparisons between participants who had seen the intervention and had not, as we wanted to compare control data to genuine ecologically valid responses as opposed to responses driven by experimenter demand. Future research would greatly further the methodology of measuring conceptual change by exploring the possibility of masking two or three session within-subject designs or multi-session measurements extending over a greater timescale.

Third, the measures used in the Main Study that were drawn from the literature provide a limited view into what changed in the minds of participants. While coding the three measures allow for clean binary signals about the presence of the content of interest, each measure only produced between one and four data points about each participant's conceptual content. This reveals only a small subset of a participant's overall conceptual content, to say nothing of the content of their larger conceptual schema. The SFP and Completion task additionally under-count the conceptual content of interest. For example, "extinct" will not be someone's only salient property of DINOSAUR, and it will not always be among a participant's responses to free-response questions measuring salient properties. Therefore while the SFP and Completion tasks are not the best measure for approximating the prevalence of specific content in a population, they are still useful to test how content changes in response to interventions. In future work, alternative question types such as ranking questions or think-aloud protocols may yield richer and more accurate measurement of conceptual content from participants.

4.2 Advancing Global and Local Implementation Questions

As argued in Section 1, conceptual engineers have not made a clear distinction between factors that exert influence on implementation at a global level (i.e., all concepts) and those that impact implementation at the local level (i.e., specific concepts), and without direct empirical data concerning the implementation of specific concepts, it will be difficult to disentangle what sort of factors affect the propagation of all concepts versus what factors are limited to specific concepts. Given the study's limitations, what tentative conclusions can therefore be drawn about global and local implementation questions?

Starting with global implementation questions, if both concepts behaved like DINOSAUR did, our results would have suggested that participants generally exhibit a high degree of resistance to conceptual revision. In such a scenario, we could have inferred that the process of implementing conceptually revised content is significantly more challenging than initially expected. This would be consistent with a particularly pessimistic reading of Machery's attractor view (Kõiv 2024; Machery 2021)—in which psychological factors naturally attract us to certain conceptual content—where attractors are the norm, not the exception. In contrast, if DINOSAUR had been readily revised, it would have suggested even relatively surprising changes could be easily propagated in adults once the effort is put in to propagate the changes.

Our findings confirm the theorizing by others (Fischer 2020; Koslow 2022; Machery 2021) that the picture is significantly more complex. Across all four interventions, we observed that individuals can be effectively instructed to categorize objects based on new classification schemes with relatively short interventions. Spending a few minutes informing individuals why Pluto is not a planet and why birds are dinosaurs brought about substantial changes in participants' classification patterns. Across all interventions, when presented with an image of Pluto and asked about its planetary status or shown a picture of a seagull and queried about its classification as a dinosaur, participants successfully categorized the objects according to the taught schema. Consequently, it appears that the process of revising individuals' mental representations in order to elicit accurate responses to questions regarding the superordinate categorization of Pluto and birds is relatively straightforward.

If the objective of a conceptual engineering project is to induce individuals to modify their explicit classification scheme, these findings are encouraging. To some engineers, this might be enough. However, the aspirations of many engineers extend beyond merely altering individuals' classifications (Isaac et al. 2022). Their ultimate goal lies in transforming people's reasoning patterns, enabling them to draw inferences based on their newfound knowledge and perceive the world through different perspectives. Accomplishing this likely necessitates the modification of individuals' associations and implicit reasoning processes.

Changing implicit associations and default information retrieval looks to be much more difficult. While the Categorization task uniformly resulted in large differences in expert-inconsistent responses between the control group and test groups, the Completion task and the Semantic Feature Production task results were mixed. In the Completion task and the Semantic Feature Production task, sizable differences were seen between the control and test groups for PLANET—a difference of around 20% expert-inconsistent responses among the conditions' participants—but no such changes were observed for DINOSAUR. Consequently, it appears that modifying individuals' explicit classification schema is significantly more feasible than altering salient features and what information is retrieved by default.

Turning now to local questions—that is, questions about what factors influence the revision of specific concepts—the contrast of different conditions and concepts offers two insights into why we were successful in revising PLANET. First, the control data in the masked time-lagged study and responses in the Follow-up Study both indicate the revision of PLANET is better known than the revision of DINOSAUR among adults. Perhaps the increased awareness of PLANET played an important role in its success, suggesting conceptual revision is something that should be built up to instead of something that can be done in a single intervention (see Carey 2011). The difference in effectiveness of the two PLANET stimuli offers another clue.[13] Adding pictures and detail to our stimuli did not by itself have an effect on conceptual revision, as evidenced by lack of revision of DINOSAUR among participants who saw the longer intervention containing images. However, length and different format appear to have played *some*

[13] Some caution is required here, as the difference between the two PLANET stimuli was smaller than 15% across all three measures.

role in uptake of the revised PLANET, suggesting that multi-modal formats may be a useful tool for propagating revisions like it.

To summarize, results support a few tentative conclusions about propagation that may serve as the basis for future work:

- **Global Level:**
  1. Short explanatory stimuli are capable of revising some natural kind concepts.
  2. For natural kind concepts, changing classification patterns appears to be relatively straightforward.
  3. For natural kind concepts, it is more difficult to modify associations and information retrieved by default than classification patterns.
  4. Changes involving splitting and/or shrinking concepts may prove easier than changes involving combining and/or expanding concepts (see Fischer 2020).
- **Local Level:**
  1. PLANET's ease of revision may have depended on previous exposure.
  2. Visual aids may have helped revise PLANET.

Up until this point, our discussions and interpretations have assumed that our measures capture conceptual revision in PLANET. In the next section, we respond to the objection that in fact the measures capture something non-conceptual.

4.3 Objection: This Is Only Belief Revision

How confident can we be that we observed a conceptual change instead of a belief-based change? The question is complicated by two factors. First, there is substantial disagreement among content internalists in philosophy and cognitive science about how to demarcate conceptual content and how content is structured (see Bloch-Mullins 2018; Quilty-Dunn 2021; Vicente and Martínez Manrique 2016; Yee and Thompson-Schill 2016), which leads to different views about where the line between beliefs and concepts lie. Answering this first complication is beyond the scope of this paper. Given the complexity of the literature on conceptual structure, the measures we used above will not satisfy everyone, nor would it be feasible to test every account in the literature at once. Nonetheless, one of the primary purposes of this paper is to provide methodological scaffolding for future projects. Anyone who believes a different measure would better capture what they think conceptual content is can easily swap in their own measures.

Second, on many content internalist frameworks, the distinction between beliefs and concepts is very subtle. This second complication is something the above experiment *does* accomplish to navigate, at least on one prominent account of content internalist conceptual engineering. The two invariantist conceptual engineers whose work we draw the Completion measure from, take concepts to be belief-like in that they are stable bodies of information (Fischer 2020; Machery 2017). On their account, what makes concepts unique is that they are retrieved quickly, automatically, and independent of context, which means they affect inferences, categorizations, and other cognitive functions (Machery 2017, 210-211). For example, someone may correctly

believe that—due to the geography of French Guyana—France's longest border is with Brazil, but not have this information elicited quickly, automatically, and independently of context when they read the word "France". Instead, the information retrieved by default may involve Paris, the French flag, France's European borders, and/or facts about France's economy. This retrieved information, as it changes from person to person, is someone's concept of France on such an account, while the belief about France's border with Brazil is not. As discussed above, the masked time-lagged design was designed to test responses to stimuli without the contextual and pragmatic salience of the intervention. Thus, the design approximates default, low-context responses—that is, content as opposed to mere belief according to Machery (2017) and Fischer (2020)—as much as is methodologically possible (for general doubts about the possibility of this, however, see Casasanto and Lupyan 2015).

Setting aside specific frameworks of conceptual content, consider some general reasons to think that the above results can be viewed as capturing conceptual revision in action. This needs to be discussed in two parts because in the above experiment, the three measures followed two patterns. The Categorization task—a multiple choice question about scientific kinds—changed dramatically, regardless of stimuli or concept. The other two tasks, Completion and SFP, both of which were open-ended free responses, only significantly changed in response to one of the four stimuli. This suggests that the Categorization task is picking up on different, more plastic, phenomena than the SFP and Completion tasks. Categorization is taken to be one of the key functions of concepts (Bloch-Mullins 2018; Machery 2009), and so we take the default reading of these findings to be related to concepts. However, one significant possibility, not ruled out here, is that the more plastic phenomena are beliefs as opposed to anything properly considered concepts. On such a reading of the results, participants are changing how they categorize Pluto and seagulls, not because of any change in conceptual content, but because of belief-level phenomena, such as the belief in the facts *Pluto is not a planet* and *birds are dinosaurs*. That is, the Categorization task, despite the intention behind its inclusion in this test, is acting as more of a test of scientific knowledge rather than how participants automatically classify objects.

However, beliefs are not the only thing that would explain the increased plasticity of the Categorization task over the other two tasks. Another possibility is that the Categorization measure is more sensitive to particular kinds of conceptual change than the SFP and Completion measure, namely the development of polysemy.[14] For example, there are multiple terms that are polysemous in that they have a loose folk meaning and a more precise scientific meaning. In the everyday sense of "fruit", tomatoes are not fruit, but in the technical sense of "fruit" tomatoes are fruit (Engelhardt 2019; Landes 2021; Machery and Seppälä 2011). Being able to understand the senses in which tomatoes are and are not fruits requires two distinct concepts, specifically

[14] Again, the complexity of the literature raises problems here, as there is deep and widespread disagreement about the relationship between content and polysemous senses (for an introduction, see Vicente 2018). We, however, take this debate to be orthogonal to our point here. Regardless of whether you think two polysemous senses are part of the same concept's content or whether you think polysemous senses are contextual in nature, the Categorization measure can be interpreted as showing the sort of difference in content that distinguishes monosemous versus polysemous words or as showing differences in response in a specific, experimentally-manipulated context.

fruit as a culinary/social kind and fruit as a botanical kind. A possible outcome of attempting to revise PLANET and DINOSAUR is that, like FRUIT, people end up with two concepts—folk and scientific counterparts to each other—which the Categorization task is for some reason more sensitive to in a way the SFP or Completion tasks are not.

That said, even if we grant that the Categorization task is merely picking up on changes of belief, this skepticism does not extend to the other two measures. For one, the SFP task does not seem to be picking up on beliefs because the information it collects is not obviously propositional and so not obviously belief-driven in the way the Categorization task may be. The SFP task asks participants to list what features come to mind related to some object. Therefore, the SFP task appears to be best interpreted as measuring salient properties of the concept as opposed to mere beliefs about the kinds (Machery 2009; McRae et al. 2005) and so is perhaps best understood as measuring stereotypical or prototypical information. Moreover, because the results of the Completion Task closely follow the results of the SFP task in all interventions, the responses in the Completion and SFP tasks appear to have a common etiology.

4.4 Expanding the Method to Other Conceptual Engineering Frameworks

The experiment and resulting data has been discussed through the lens of content internalist conceptual engineering, which understands the target of conceptual engineering to be concepts and understands concepts to be token psychological entities. While this is currently a popular conceptual engineering framework, it is by no means the only one on the market. Other frameworks include those that propose that the goal of conceptual engineering is semantic meaning (Cappelen 2018; Sterken 2020), speaker meaning (Pinder 2020, 2021), or conceptual content that is grounded in facts external to individuals (Haslanger 2020; Sawyer 2020; Scharp 2013). In this section, we discuss the significance of the masked time-lagged method to other frameworks. We specifically focus on two prominent families of frameworks that focus on language instead of concepts, namely speaker meaning accounts and semantic externalist accounts.

The methods described above can easily be expanded to speaker meaning accounts of conceptual engineering. Speaker meaning accounts of conceptual engineering take the goal of conceptual engineering to be to change what people take themselves to mean by the words that they utter (Pinder 2020, 2021). Speaker meaning conceptual engineers do not aim to change what a word means in some broad, interpersonal sense. Instead they target the intentions and beliefs speakers have related to using a specific term (although on some frameworks, linguistic intentions, linguistic beliefs, and meaning go together). Thus, similar to content internalist frameworks, speaker meaning conceptual engineering places the target of conceptual engineering in token psychological states. The best way to test what speakers intend to mean by a term is to have them produce speech acts using the term and coding how the term is used. For that reason, the Completion task may be a suitable measure, although multi-sentence productions would provide richer sources of data about speaker intentions.

When it comes to applying the findings of our studies to semantic externalist frameworks of conceptual engineering, the extension of this study is not as straightforward because semantic externalists do not place the focus of their theories on token psychological states. Nonetheless, when we dig into the role token psychological states generally play in semantic externalist accounts of meaning, we can see that not only can masked time-lagged designs answer questions about how to spread true beliefs about meaning or reference changes that has already occurred, it can also answer questions about how best to use our limited ability to change externalist meaning or reference.

Externalists will not want to say that our experiment revised the concept PLANET or meaning of "planet". They will instead contend either that a) "planet" always excluded Pluto (e.g., Ball 2020; Kripke 1980, b) the concept or meaning was revised in 2006 by the International Astronomical Union, or c) revision of "planet" or PLANET occurred at some later date when, for example, the new linguistic norm became the dominant one among English speakers (Evans 1973).[15] Even if semantic externalists contend our experimental data does not reveal anything about how to change meaning, the experiment does shed light on how conceptual engineers can help individuals become aware of changes in meaning. While on many of these frameworks, the meaning of people's utterances containing"planet" has changed since 2006, these sorts of changes can happen without people's awareness (see Pollock 2021; Wikforss 2015). Thus, externalists can still view the experiment as testing phenomena related to propagation. While the experiment did not change meaning, it propagated true beliefs about the meaning of "planet" in light of changes that have occurred.

The masked time-lagged design can do more than show how to spread true semantic beliefs related to revisions, however. Semantic externalists can use the masked time-lagged design proposed here to determine the most effective ways to change meaning—even if meaning is grounded externally to individual speakers. To see this, we need to focus on what externalists take ground semantic facts. For semantic externalists, the meaning of a word-type or utterance-type (that is, what a word means, in general) is determined by some combination of linguistic facts and non-linguistic facts. While they disagree about which linguistic and non-linguistic facts matter, many, if not most non-linguistic facts, such as the joints of natural kinds, are outside of conceptual engineer controls (Cappelen 2018). This limits the scope of what sort of changes are possible. Nonetheless, many meaning-determining linguistic facts are within the scope of empirical methods (Koslow 2022; Sterken 2020; Thomasson 2021), and a subset of those meaning-determining facts are things conceptual engineers have (limited) influence over. This is because many linguistic facts are grounded in, among other things, intentions and beliefs of individual speakers—that is, token psychological states (see Nimtz 2024).

To illustrate how global and local questions about intentional semantic externalist meaning change can be tested, consider a version of an Evans-style metasemantics, where the reference of a word-token is determined by the dominant causal source of all the uses of that word-token (Evans 1973; Leckie and Williams 2019). On this view, my

[15] That said, the data in the control group raises doubts that it has in fact become the dominant use of "planet".

use of "cat" refers to cats because most of the people around me use "cat" in a way that traces back to cats. What things there are in the world to be a dominant causal source is by and large outside of conceptual engineers' control. Nonetheless, this Evan-style view locates part of the ground of reference in the collective uses of a community of speakers. The collective use of a community of speakers is, at rock bottom, a large number of token beliefs and practices about language. Therefore, changing meaning by changing psychological states is possible if enough psychological states change to disrupt the current dominant source (see Sterken 2020). How to best use resources to disrupt the dominant source is something experimental methods can study. In fact, this will look similar to the speaker meaning account discussed above, as it will involve studying how people shift the way they speak in light of an intervention.[16]

The Evans-style view is merely illustrative of the larger constellation of semantic externalist theories. Semantic externalists generally take some sort of linguistic norm or practice to play a role in determining the meaning of word-types, such as the collective use of a causal historical chain (Kripke 1980) or linguistic conventions (Lewis 1969). Linguistic norms and practices are partially or wholly composed of linguistic beliefs and intentions, although the way linguistic beliefs and intentions combine to form norms may be extremely complicated (Lewis 1969; Nimtz 2024). This means that semantic externalists can study how to change meaning by studying how to bring about collective shifts in linguistic beliefs and intentions. Here we find ourselves in the realm of token psychological entities, and so semantic externalists can test what factors influence meaning or reference change using variations of the above empirical masked time-lagged design. Granted, no amount of experiments will make externalists semantically omnipotent—they will still be limited in what they can change by linguistic and non-linguistic facts outside of their control. This is not a problem unique to externalism, however. As discussed above, internalist conceptual engineers will also be limited by features of our cognition that will prevent certain proposed changes from taking hold. Nonetheless, the more we learn through experiments about how the grounds of conceptual content or semantic facts change, the better conceptual engineers will be at revising or replacing conceptual content or semantic facts.

5 Conclusion

Many philosophers have recently begun theorizing that the aim of philosophy should be to develop revised concepts and spread those revised concepts to the right people. Little is understood about how the propagation of such revised concepts could be spread or even measured. In this project, we demonstrated how a masked time-lagged design could be used to directly test the revision of participants' concepts. We specifically attempted to revise DINOSAUR and PLANET, finding tantalizingly mixed results. While our stimuli appear to have successfully revised PLANET in participants, the same cannot be straightforwardly said for DINOSAUR. Therefore, further work is needed to study

[16] This is assuming the framework is productivist. An interpretationalist semantic externalist could study how interpretations of statements change after intervention. See Simchen (2017).

the process of propagating conceptual changes for the ends of conceptual engineering now that an empirical framework is in place.

Author Contributions Ethan Landes: Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing; Kevin Reuter: Conceptualization, Data curation, Funding Acquisition, Methodology, Project Administration, Writing - original draft, Writing - review & editing.

Funding Project supported by Swiss National Science Foundation grant 181082.

Data Availability We report all of our key measures, manipulations, and exclusions, and all data and experiment materials are available for download at the Open Science Framework: https://osf.io/vjxsn/.

Declarations

Conflicts of Interest The authors have no conflicts of interest to report.

References

Ambridge, B., J. M. Pine, C. F. Rowland, F. Chang, and A. Bidgood. 2013. The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science* 4 (1): 47–62.

Andow, J. 2020. Fully experimental conceptual engineering. *Inquiry* 1–27. https://doi.org/10.1080/0020174X.2020.1850339.

Andreasen, N.C. 2010. Posttraumatic stress disorder: A history and a critique. *Annals of the New York Academy of Sciences* 1208 (1): 67–71. https://doi.org/10.1111/j.1749-6632.2010.05699.x.

Ball, D. 2020. Relativism, metasemantics, and the future. *Inquiry* 63 (9–10): 1036–1086. https://doi.org/10.1080/0020174X.2020.1805710.

Barsalou, L.W. 1983. Ad hoc categories. *Memory & cognition* 11:211–227.

Belleri, D. 2021. On pluralism and conceptual engineering: Introduction and overview. *Inquiry* 1–19. https://doi.org/10.1080/0020174X.2021.1983457.

Bird, A., and E. Tobin. 2023. Natural Kinds. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta, and U. Nodelman. Spring 2023 ed. Metaphysics Research Lab, Stanford University.

Bloch-Mullins, C.L. 2018. Bridging the Gap between Similarity and Causality: An Integrated Approach to Concepts. *The British Journal for the Philosophy of Science* 69 (3): 605–632. https://doi.org/10.1093/bjps/axw039.

Brusatte, S. 2017. *How Birds Evolved from Dinosaurs*. https://www.scientificamerican.com/article/how-birds-evolved-from-dinosaurs/. https://doi.org/10.1038/scientificamerican0117-48.

Buskell, A. 2017. What are cultural attractors? *Biology & Philosophy* 32 (3): 377–394. https://doi.org/10.1007/s10539-017-9570-6.

Cappelen, H. 2018. *Fixing Language*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198814719.001.0001.

Carey, S. 2011. *The origin of concepts*, 1st ed. Oxford: Oxford Univ. Press.

Casasanto, D., and G. Lupyan. 2015. All Concepts Are Ad Hoc Concepts. In *The Conceptual Mind*, ed. E. Margolis, and S. Laurence, 543–566. The MIT Press. https://doi.org/10.7551/mitpress/9383.003.0031.

Chang, K. 2022. Is Pluto a Planet? What's a Planet, Anyway? *The New York Times*.

Conrad, F. G., M. F. Schober, and N. Schwarz. 2014. Pragmatic Processes in Survey Interviewing. In *The Oxford Handbook of Language and Social Psychology*, ed. T. M. Holtgraves, 0. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199838639.013.005.

Cullen, S. 2010. Survey-Driven Romanticism. *Review of Philosophy and Psychology* 1 (2): 275–296. https://doi.org/10.1007/s13164-009-0016-1.

Egré, P., and C. O'Madagain. 2019. Concept Utility. *The Journal of Philosophy* 116 (10): 525–554. https://doi.org/10.5840/jphil20191161034.

Eklund, M. 2011. What are Thick Concepts? *Canadian Journal of Philosophy* 41 (1): 25–49. https://doi.org/10.1353/cjp.2011.0007.

Engelhardt, J. 2019. Linguistic labor and its division. *Philosophical Studies* 176 (7): 1855–1871. https://doi.org/10.1007/s11098-018-1099-2.

Evans, G. 1973. The Causal Theory of Names. *Aristotelian Society Supplementary* 47 (1): 187–225.

Fischer, E. 2020. Conceptual control: On the feasibility of conceptual engineering. *Inquiry* 1–29. https://doi.org/10.1080/0020174X.2020.1773309.

Fischer, E., and P. E. Engelhardt. 2019. Lingering stereotypes: Salience bias in philosophical argument. *Mind & Language* mila.12249. https://doi.org/10.1111/mila.12249.

Fischer, E., P. E. Engelhardt, and A. Herbelot. 2015. Intuitions and illusions: From explanation and experiment to assessment. In *Experimental Philosophy, Rationalism, and Naturalism. Rethinking Philosophical Method*, ed. E. Fischer, and J. Collins, 259–292. Routledge.

Gelman, S.A. 2013. Artifacts and Essentialism. *Review of Philosophy and Psychology* 4 (3): 449–463. https://doi.org/10.1007/s13164-013-0142-7.

Gelman, S. A., and E. M. Markman. 1987. Young Children's Inductions from Natural Kinds: The Role of Categories and Appearances. *Child Development* 58 (6): 1532. https://doi.org/10.2307/1130693.

Giora, R. 2003. *On our mind: Salience, context, and figurative language*. Oxford: Oxford University Press.

Griffiths, P. E., and E. Machery. 2008. Innateness, Canalization, and 'Biologicizing the Mind'. *Philosophical Psychology* 21 (3): 397–414. https://doi.org/10.1080/09515080802201146.

Hampton, J.A. 1979. Polymorphous concepts in semantic memory. *Journal of verbal learning and verbal behavior* 18 (4): 441–461.

Haslanger, S. 2000. Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs* 34 (1): 31–55. https://doi.org/10.1111/0029-4624.00201.

Haslanger, S. 2020. How Not to Change the Subject. In *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability*, ed. T. Marques, and Å. Wikforss, Oxford University Press.

IAU. 2006. *Definition of a Planet in the Solar System* (Tech. Rep. No. Resolution B5). Paris: International Astronomical Union.

Isaac, M.G. 2023. Which Concept of Concept for Conceptual Engineering? *Erkenntnis* 88 (5): 2145–2169. https://doi.org/10.1007/s10670-021-00447-0.

Isaac, M. G., S. Koch, and R. Nefdt. 2022. Conceptual engineering: A road map to practice. *Philosophy Compass* 17 (10): e12879. https://doi.org/10.1111/phc3.12879.

Jorem, S. 2021. Conceptual Engineering and the Implementation Problem. *Inquiry: An Interdisciplinary Journal of Philosophy* 64 (1-2): 186–211. https://doi.org/10.1080/0020174x.2020.1809514

Kitsik, E. 2023. Epistemic paternalism via conceptual engineering. *Journal of the American Philosophical Association* 9 (4): 616–635. https://doi.org/10.1017/apa.2022.22.

Knobe, J., S. Prasada, and G. E. Newman. 2013. Dual character concepts and the normative dimension of conceptual representation. *Cognition* 127 (2): 242–257. https://doi.org/10.1016/j.cognition.2013.01.005.

Koch, S. 2021. The externalist challenge to conceptual engineering. *Synthese* 198:327–348. https://doi.org/10.1007/s11229-018-02007-6.

Kõiv, R. 2024. To reform or to eliminate an attractor? *Synthese* 204 (2): 52. https://doi.org/10.1007/s11229-024-04685-x.

Kornblith, H. 1997. *Inductive interference and its natural ground: An essay in naturalistic epistemology*. Cambridge, Mass.: MIT Press.

Koslow, A. 2022. Meaning change and changing meaning. *Synthese* 200 (2): 94. https://doi.org/10.1007/s11229-022-03563-8.

Kripke, S. 1980. *Naming and Necessity*. Oxford, UK ; Cambridge, USA: Blackwell Publishers.
Landes, E. 2021. *Philosophy and philosophy: The subject matter and the discipline* (Thesis, The University of St Andrews). https://doi.org/10.17630/sta/1072
Landes, E. 2025. Conceptual Engineering Should be Empirical. *Erkenntnis*. https://doi.org/10.1007/s10670-025-00923-x.
Landes, E. in press. How Language Teaches and Misleads: "Coronavirus" and "Social Distancing" as Case Studies. In *New Perspectives on Conceptual Engineering*, ed. M. G. Isaac, S. Koch, and K. Scharp. Synthese Library.
Leckie, G., and R. Williams. 2019. Words by convention. In *Oxford Studies in Philosophy of Language*, ed. E. Lepore, and D. Sosa, 1 (1). Oxford University Press.
Lewis, D.K. 1969. *Convention: A philosophical study*. Cambridge, Mass: Harvard University Press.
Liao, S., and N. Hansen. 2023. 'Extremely Racist' and 'Incredibly Sexist': An Empirical Response to the Charge of Conceptual Inflation. *Journal of the American Philosophical Association* 9 (1): 72–94. https://doi.org/10.1017/apa.2021.46.
Machery, E. 2009. *Doing without Concepts*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195306880.001.0001.
Machery, E. 2017. *Philosophy within its proper bounds*, 1st ed. Oxford, United Kingdom: Oxford University Press.
Machery, E. 2021. A new challenge to conceptual engineering. *Inquiry* 1–24. https://doi.org/10.1080/0020174X.2021.1967190.
Machery, E., P. Griffiths, S. Linquist, and K. Stotz. 2019. Scientists' Concepts of Innateness: Evolution or Attraction? In *Advances in Experimental Philosophy of Science*, eds. R. Samuels, and D. A. Wilkenfeld, 172–201. Bloomsbury.
Machery, E., and S. Seppälä. 2011. Against hybrid theories of concepts. *Anthropology and Philosophy* 10:99–126.
Margolis, E., and S. Laurence. 2007. The Ontology of Concepts-Abstract Objects or Mental Representations? *Noûs* 41 (4): 561–593. https://doi.org/10.1111/j.1468-0068.2007.00663.x.
McRae, K., G. S. Cree, M. S. Seidenberg, and C. Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* 37 (4): 547–559. https://doi.org/10.3758/BF03192726.
Muehlenhard, C. L., and Z. D. Peterson. 2011. Distinguishing Between Sex and Gender: History, Current Conceptualizations, and Implications. *Sex Roles* 64 (11): 791–803. https://doi.org/10.1007/s11199-011-9932-5.
Napolitano, M. G., and K. Reuter. 2023. What is a Conspiracy Theory? *Erkenntnis* 88 (5): 2035–2062. https://doi.org/10.1007/s10670-021-00441-6.
NASA. 2023. What is a Planet? https://science.nasa.gov/solar-system/planets/whatis-a-planet/
Nimtz, C. 2024. Engineering concepts by engineering social norms: Solving the implementation challenge. *Inquiry* 67 (6): 1716–1743. https://doi.org/10.1080/0020174X.2021.1956368.
Pinder, M. 2017. Does Experimental Philosophy Have a Role to Play in Carnapian Explication? *Ratio* 30 (4): 443–461. https://doi.org/10.1111/rati.12164.
Pinder, M. 2020. Conceptual engineering, speaker-meaning and philosophy. *Inquiry* 1–15. https://doi.org/10.1080/0020174X.2020.1853342.
Pinder, M. 2021. Conceptual Engineering, Metasemantic Externalism and Speaker-Meaning. *Mind* 130 (517): 141–163. https://doi.org/10.1093/mind/fzz069.
Poling, D. A., and E. M. Evans. 2004. Are dinosaurs the rule or the exception? Developing concepts of death and extinction. *Cognitive Development* 19:363–383. https://doi.org/10.1016/j.cogdev.2004.04.001.
Pollock, J. 2021. Content internalism and conceptual engineering. *Synthese* 198 (12): 11587–11605. https://doi.org/10.1007/s11229-020-02815-9.
Quilty-Dunn, J. 2021. Polysemy and thought: Toward a generative theory of concepts. *Mind & Language* 36 (1): 158–185. https://doi.org/10.1111/mila.12328.
Quine, W.V. 1969. Natural kinds. In *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday*, 5–23. Springer.
Rescorla, L.A. 1980. Overextension in early language development. *Journal of Child Language* 7 (2): 321–335. https://doi.org/10.1017/S0305000900002658.
Reuter, K. 2019. Dual character concepts. Philosophy *Compass* 14 (1): e12557. https://doi.org/10.1111/phc3.12557.

Reuter, K. 2024. Salient semantics. *Synthese* 204 (2): 39.
Reuter, K., and G. Brun. 2022. Empirical Studies on Truth and the Project of Re-engineering Truth. *Pacific Philosophical Quarterly* 103 (3): 493–517. https://doi.org/10.1111/papq.12370.
Rose, D., and S. Nichols. 2019. Teleological Essentialism. *Cognitive Science* 43 (4): e12725. https://doi.org/10.1111/cogs.12725.
Saigh, P. A., and J. D. Bremner. 1999. The history of posttraumatic stress disorder. In *Posttraumatic stress disorder: A comprehensive text*, 1–17. Needham Heights, MA, US: Allyn & Bacon.
Sawyer, S. 2020. Truth and objectivity in conceptual engineering. *Inquiry* 63 (9–10): 1001–1022. https://doi.org/10.1080/0020174X.2020.1805708.
Sawyer, S. in press. Concepts in Conceptual Engineering. In *A Philosophical History of the Concept*, ed. S. Schmid, and H. Taieb. Cambridge University Press.
Scharp, K. 2013. *Replacing Truth*. Oxford, New York: Oxford University Press.
Schwarz, N. 1995. What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation. *International Statistical Review / Revue Internationale de Statistique* 63 (2): 153–168. https://doi.org/10.2307/1403610.
Scott-Phillips, T., S. Blancke, and C. Heintz. 2018. Four misunderstandings about cultural attraction. *Evolutionary Anthropology: Issues, News, and Reviews* 27 (4): 162–173. https://doi.org/10.1002/evan.21716.
Shields, M. 2021. Conceptual domination. *Synthese* 199 (5–6): 15043–15067. https://doi.org/10.1007/s11229-021-03454-4.
Shields, M. 2023. Conceptual Engineering, Conceptual Domination, and the Case of Conspiracy Theories. *Social Epistemology* 37 (4): 464–480. https://doi.org/10.1080/02691728.2023.2172696.
Shtulman, A., and P. Calabi. 2013. Tuition vs. intuition: Effects of instruction on naive theories of evolution. *Merrill-Palmer Quarterly* 59 (2): 141–167.
Shtulman, A., and J. Valcarcel. 2012. Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition* 124: 209–215.
Simchen, O. 2017. *Semantics, metasemantics, aboutness (First*. Edition. Oxford: Oxford University Press.
Spelke, E. S., and S. Tsivkin. 2001. Initial knowledge and conceptual change: Space and number. In *Language Acquisition and Conceptual Development*, ed. M. Bowerman, and S. Levinson, 70–98. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620669.005.
Sperber, D. 1996. *Explaining culture: A naturalistic approach*. Oxford: Oxford Blackwell.
Sterken, R.K. 2020. Linguistic Interventions and Transformative Communicative Disruption. In *Conceptual Engineering and Conceptual Ethics*, ed. H. Cappelen, D. Plunkett, and A. Burgess. Oxford University Press.
Thomasson, A. 2021. Conceptual engineering: When do we need it? How can we do it? *Inquiry* 1–26. https://doi.org/10.1080/0020174X.2021.2000118.
Vicente, A. 2018. Polysemy and word meaning: An account of lexical meaning for different kinds of content words. *Philosophical Studies* 175 (4): 947–968. https://doi.org/10.1007/s11098-017-0900-y.
Vicente, A., and F. Martínez Manrique. 2016. The Big Concepts Paper: A Defence of Hybridism. *The British Journal for the Philosophy of Science* 67 (1): 59–88. https://doi.org/10.1093/bjps/axu022.
Wakil, S. 2023. Experimental Explications for Conceptual Engineering. *Erkenntnis* 88 (4): 1509–1531. https://doi.org/10.1007/s10670-021-00413-w.
Wikforss, Å. 2015. The insignificance of transparency. In *Externalism, Self-Knowledge, and Skepticism: New Essays*, ed. S. C. Goldberg, 142–164. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107478152.009.
Willemsen, P., and K. Reuter. 2021. Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy* 10 (2): 135–146. https://doi.org/10.1002/tht3.488
Yee, E., and S. L. Thompson-Schill. 2016. Putting concepts into context. *Psychonomic Bulletin & Review* 23 (4): 1015–1027. https://doi.org/10.3758/s13423-015-0948-7.

Authors and Affiliations

Ethan Landes[1] · **Kevin Reuter**[2,3]

✉ Ethan Landes
E.Landes@Kent.ac.uk

Kevin Reuter
kevin.reuter@uzh.ch

1 University of Kent, Canterbury, United Kingdom

2 Institute of Philosophy, University of Bern, Bern, Switzerland

3 Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Gothenburg, Sweden