Journal of Memory and Language 95 (2017) 68–77

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Revising deference: Intuitive beliefs about category structure constrain expert deference

Alexander Noyes *, Frank C. Keil

Yale University, United States

ARTICLE INFO

Article history:
Received 18 October 2016
revision received 13 February 2017
Available online 1 March 2017

Keywords:
Categorization
Natural kinds
Experts

ABSTRACT

Concepts are grounded in intuitive theories, yet intuitive theories are often sparse and incomplete. Deferring to experts can potentially fill those gaps. Sometimes experts convey new information, such as discovering a new planet (Experiment 1 and 3). Other times they revise past conclusions, such as concluding that Pluto is actually not a planet (Experiment 2 and 3). For non-experts to maintain scientific accuracy, they need to assimilate the expert judgments in either case. However, we find that people are less likely to defer after revision than novel discovery. In each case, their essentialist intuitions explain the pattern of results. The more participants construe categories in essentialist terms, the more they reject category revision; the opposite occurs for novel discoveries. Moreover, people only reject revision when it conflicts with essentialist intuitions (Experiment 4). Thus, the same intuitive theories that encourage deference also constrain it.

Introduction

Words like "leopard" and "gold" are commonplace in language, but the knowledge of how to distinguish gold from other substances (like fool's gold) and leopards from other animals (like jaguars) is relatively rare. Fortunately, this knowledge is available in the broader linguistic community (Putnam, 1973). Certain experts do know gold's constitutive properties (composed of the element with 79 protons) and how to determine the presence of these properties (chemical assays). In principle, non-experts can tap into this knowledge by deferring to those who know more, enabling the greater community to self-correct and apply concepts accurately. Nevertheless, there are persistent and widespread ambiguities in the application of natural concepts (Dupré, 1981). For example, the common usage of "lily" does not align with any biological taxon. People include flowers from diverse genera but exclude plants in the same family (like tulips and onions). A similar pattern exists for words as common as reptile, fish, fruit, and butterfly. Thus, despite the availability of expert knowledge, there seem to be barriers preventing expert judgments from readily being accepted into the broader public. Here, we propose that the very same cognitive mechanisms that enable deference also constrain it.

Concepts are often said to be embedded in people's causal beliefs or intuitive theories and to derive many of their properties from those causal-theoretical frameworks (Medin & Ortony, 1989; Medin, Wattenmaker, & Hampson, 1987; Murphy, 2002; Murphy & Medin, 1985). These causal beliefs provide an organizing and cohering function that allows people to make sense of category-linked properties and their relationship to each other. In turn, these causal beliefs drive category learning and category judgments. Yet, people's causal beliefs are often sparse and incomplete (Keil, 2003, 2010) – participants represent categories through placeholder essences and skeletal fragments rather than full-fledged understandings of the underlying causal relations. For example, many people understand that jaguars and leopards are different species of felines. Probably very few people, however, could describe what it is about their anatomy, genetic structure, or ancestry that explains why they belong to distinct species. These incomplete representations are supplemented by the availability of experts and expert knowledge. More specifically, even though laypeople's causal theories are insufficiently detailed to categorize many entities in the world, their intuitive theories may guide them to find the relevant experts that *can* determine category membership (Keil, 2003, 2010). Thus, intuitive theories may help people outsource more difficult judgments to knowledgeable experts. To fully understand the relationship between intuitive theories and concepts, then, requires understanding how people interact with

* Corresponding author at: Department of Psychology, Yale University, Box 208205 [Courier: 2 Hillhouse Ave], New Haven, CT 06520-8205, United States.

E-mail address: Alexander.Noyes@yale.edu (A. Noyes).

http://dx.doi.org/10.1016/j.jml.2017.02.003
0749-596X/

outside sources of information (i.e., experts) through the mechanism of deference.

Deference is also central to psychological and philosophical theories of semantic externalism – the idea that word meaning (specifically, a concept's referent) is determined by the reality of the world rather than an individual's beliefs (Jylkkä, 2008; Kripke, 1972; Margolis, 1998; Millikan, 1998; Putnam, 1973). From this position, the meaning of "gold" is linked to the underlying nature of gold, a nature that only experts may be able to detect. In that sense, the meaning of gold is external – outside of the head of individual speakers. This position is most consistently defended for natural kinds, including taxonomic biological kinds (e.g., tigers and elm trees) and chemical substances (e.g., gold and water), where objective criteria are the most plausible (Putnam, 1973; Schwartz, 1978, 1979; Sloman & Malt, 2003).

More directly relevant to psychological theory, however, is whether people's own intuitions conform to externalism. In support of this proposal, people treat natural kinds as if their membership is objectively discoverable (Diesendruck & Gelman, 1999; Keil, 1989; Rips, 1989; cf. Kalish, 2002). For example, in their use of linguistic hedges (Malt, 1990) people apply "according to experts, this is an X" selectively to natural kinds (reflecting their presumed objectivity) and are more reluctant to apply the hedge "loosely speaking, this is an X" (reflecting that natural kinds have clear-cut boundaries). In contrast, the opposite pattern of acceptability judgments is found for artifacts (suggesting that people do not believe they possess necessary and sufficient features, Malt & Johnson, 1992 cf. Bloom, 1998). Furthermore, people believe that ambiguous animals like an apparent tiger-lion must be either a tiger or a lion and that experts can correctly determine to which category the animal belongs (Coley & Luhmann, 2000). Finally, participants reliably express externalist intuitions when reasoning about scientific discovery – believing a newly discovered kind is categorically distinct when it has a novel underlying structure even if it resembles known kinds superficially (Jylkkä, Railo, & Haukioja, 2009; cf. Braisby, Franks, & Hampton, 1996). The dominant explanation for this pattern of findings is psychological essentialism (Gelman, 2003): the belief that certain types of categories are based in underlying causal properties – essences – that endow them with category-typical properties. To illustrate, if one believes that water is based in H_2O, then a liquid is water not because it appears to be (or has water-like properties) but because it is composed of H_2O; therefore, water is objective (by being determined by real-world properties) and has clear boundaries (is composed of H_2O or not).

Externalism, however, is not restricted to strict versions of psychological essentialism. It only requires that participants are sensitive to real-world causal relations (Hampton, Estes, & Simmons, 2007; Malt, 1994; Patalano, Chin-Parker, & Ross, 2006; Strevens, 2000; cf., Ahn et al., 2001). The literature demonstrates that there are clearly domains of categories, such as categories of animals and substances, that people believe reflect the real structure of the world (Gelman, 2003), even though their own understandings of the real world are simplified and incomplete (Keil, 2003). Thus, participants' reliance on external information (and thus the role of experts) is not radically altered by replacing strict essentialism with alternative proposals that emphasize causal relations (such as Hampton et al., 2007, or Strevens, 2000).

Despite these attempts to maintain external validity, errors in categorization and deference abound; inaccuracies often stem from the same cognitive mechanisms that support external validity. For example, participants have surprisingly similar epistemic relationships (Keil, 1989). The underlying nature of things is complicated even in the case of chemistry, which has the closest approximation to single property essences (e.g., water = H_2O); in general, properties do not result from single causal forces but from the complex and mutually reinforcing interactions of many causal forces (Boyd, 1999).

Withdrawing deference in the context of revision is not necessarily irrational. Nor is this skepticism specific to deferential concepts. Rather, it likely stems from more general commonsense notions about objective and absolute judgments – for example, one uses a litmus test as a decisive test of whether a liquid is an acid or a base. If the litmus test produced a different answer the second time one dipped it into the liquid, it would imply the strip was faulty (neither judgment could be considered decisive). Categorization of natural kinds may often be viewed as far more like a litmus test than it really is. Thus, this normally rational reasoning process may become problematically generalized to the greater scientific process and to expert category judgments.

Thus, the primary error is one of calibration – inferring from oversimplified or essentialist beliefs that expert categorization is as decisive as other objective tests (like a litmus test). To the extent that natural kinds are objective categories, they are also complex and conceptually difficult to pin down (Boyd, 1999; Dupré, 1981; tively). This similarity likely stems from the overlaps between how both types of beliefs are acquired (Lane & Harris, 2014) – namely, through deference to others deemed confident and trustworthy. Furthermore, essentialist intuitions (such as belief in stable and immutable categories, e.g., Gelman, 2003) are a major obstacle to the accurate understanding of evolution (Gelman & Rhodes, 2012; Shtulman, 2006; Shtulman & Schulz, 2008). Across the lifespan, intuitive theories seem to co-exist and interfere with scientific knowledge (Bloom & Weisberg, 2007; Shtulman & Harrington, 2016; Shtulman & Valcarcel, 2012).

These findings ironically suggest that though people believe their concepts correspond to the real world (and attempt to maintain external validity), their concepts are nevertheless plagued by errors. Our studies examine this tension, rethinking past accounts of deference and adding to the understanding of how intuitive theories constrain conceptual change and scientific accuracy. Indeed, we aim to understand why people are so often inaccurate even for the most basic natural kinds (Dupré, 1981). Category revision is common in science, occurring when experts conclude that an entity belongs to a different category than was previously assumed. For example, category revision occurred when experts concluded that Pluto was a dwarf planet rather than a planet. We propose that one mechanism by which people become error prone is the rejection of category revision. Past accounts often focused on novel classifications, such as experts concluding that a newly discovered astronomical body is a star. Yet, by the nature of the scientific process, sometimes experts change category judgments they had previously made, as occurred very publicly in the case of Pluto. Past accounts do not distinguish between novel categorization and category revision – even though both types of expert judgments are equally important parts of the scientific process.

Category revision may challenge many people's beliefs about the underlying nature of categories and thus the implications of an expert's category judgment. People may believe that the underlying natures of things are more straightforward than they really are (Marsh & Rothman, 2013). They may not understand that essences (if there are any, Dupré, 1981; Leslie, 2013) are substantially more complicated than single properties and linear causal

stances towards science and the supernatural (Shtulman, 2013). Thus, the most common justification participants provided for their beliefs about both science and the supernatural was their trust in authorities (scientific and religious authorities, respec-

Leslie, 2013). Past accounts of deference have not considered how people's beliefs about natural kinds (as objective and having clear boundaries) may ironically tap into more general reasoning about objective and absolute judgments – and thus may lead

participants towards withdrawing their reliance on experts. Thus, we propose that the same cognitive underpinnings that encourage deference to experts about category membership (e.g., the belief that category membership is objective and clear cut) may also inhibit expert deference, leading to conceptual inertia and inaccuracy.

To test this proposal we provided subjects with two types of expert judgments: novel classification and category revision. We then measured whether participants deferred to expert judgments. By using science-relevant categories that naturally ranged in how "essentialist" they were (how meaningful, objective, and clear-cut their boundaries were) we were able to test the relationship of essentialist-type intuitions and deference. In the context of novel classification we expected that deference would be high overall and that deference would be highest for the most essentialist categories – fitting with previous literature and the classic account of deference. In the context of category revision, however, we expected that deference would be substantially lower and would be *lowest* for the most essentialist categories – reflecting that people's essentialist intuitions lead them to be skeptical of category revision.

In short, the essentialist intuition that natural kinds possess features comparable to classic definitions may lead to the false impression that once scientists lock on to those features there can be little allowance for later revisions. When scientists do revise, the entire deference process becomes undermined for the layperson.

Experiment 1

We first attempted to examine basic patterns of deference. In the case of novel discovery (when an expert classifies an object for the first time) participants should be very willing to accept the expert's category judgments. Furthermore, participants should be most willing to defer when the category boundaries are construed in the most highly essentialist terms. To test this we had participants respond to a variety of natural terms on essentialism measures and a deference measure. We choose not to highlight the deference prompt as separate from the other category measures in order to minimize task demands.

We presented participants with a large variety of natural terms, the breadth of which was larger than is typically considered in prior work (which tends to focus on biological kinds). This breadth helps ensure the generalizability of our findings. Items varied on whether they occurred at the basic level or a relatively superordinate or subordinate category level, as essentialism can occur at all such levels (Gelman, 2003). Furthermore, although all of the terms were natural and scientific, they varied on whether they might count as natural kinds (as defined in the literature; e.g., Schwartz, 1979). For example, tiger is one of the oldest examples proposed to be a natural kind (e.g., Putnam, 1973), but meteoroid is most certainly not a natural kind term. We suspected participants would view most the terms as examples of natural kinds on average but would also be sensitive to this variation.

We had participants' rate category boundaries rather than the categories themselves. We chose this design for three reasons: First, boundaries make the task more concrete for participants. For example, it is unclear how to evaluate how clear-cut or graded a boundary of a single category is without specifying the boundary to these experiments is the contrast between X and Y. Thus, using these same contrasts in Experiment 1 maximizes overlap between experiments. Third, categorical contrasts provide more variability on measures of essentialism than single categories. For example, participants likely construe the majority of natural categories (e.g., "planet") in essentialist terms (objectivity, meaningful boundaries, etc.), but by providing multiple contrasts we can introduce more variation: For example, the boundary between the categories of planet and dwarf planet is likely viewed as being more graded and less objective than the boundary between the categories of planet and star. We employed this methodological decision throughout all of the reported experiments.

Method

Participants

We aimed for approximately 50 participants on the expectation of medium effect sizes. Fifty-eight participants were recruited from Amazon Mechanical Turk who resided in the US and who had a 95% approval rating. Participants were paid $1.50 for completion of the experiment. The median education level was a 4-year degree, which 47% of participants had. 38% of participants had completed some college, 8% had only high school degrees, and 10% had graduate or professional degrees.

Design and procedure

The design of the experiment was to explore how participants' essentialist intuitions about natural terms affected their willingness to defer to experts. All essentialism and deference-related measures were blocked together by item, such that for any given natural term participants answered all measures before moving on. Within a block, all measures were randomized. Blocks were also presented in randomized order.

Participants were instructed that they would rate 15 categorical distinctions (the natural terms, described below) on 5 dimensions (the rating scales and deference prompt, described below). Participants were instructed not to consult outside sources. After responding to all 15 natural terms, participants answered questions about their educational attainment.

Stimuli. Participants rated scientific terms from the following domains: astronomy (planet-dwarf planet, gas giant-solid planet, planet-star, meteor-meteoroid), animal and plant biology (tiger-lion, reptile-bird, dog-wolf, tree-bush, plant-algae, lily-tulip, turtle-tortoise, animal-plant), and chemistry (ruby-sapphire, diamond-graphite, solid-liquid). Participants rated the boundary between two categories (e.g., planet-star) rather than any individual category (e.g., planet).

Ratings. Participants rated the category boundaries on four 7-point Likert scales (Table 1) where only the end points were marked. These measures tapped into four aspects of essentialism (and externalism) (Gelman, 2003): subjective vs. objective, graded vs. bounded, inductively weak vs. inductively meaningful, and invented/legislated vs. discoverable (see Table 1). For every trial, participants had the entire prompt described in Table 1 at their disposal so that they could always re-read the instructions. The last sentence was bolded, however, so that participants could skip the instructions after they understood the scales. Overall, these

condition; the boundary between a planet and a dwarf planet is quite different than the boundary between a planet and a star. By providing the contrast it should be clearer to participants how to interpret the question. Second, it helps match Experiment 1 to the subsequent experiments exploring contexts of category revision. In those experiments an object is originally categorized as an X (e.g., bird) and is re-categorized as a Y (e.g., reptile); intrinsic

measures explored how closely people believe these categorical boundaries correspond to the real world.

We measured deference by asking participants how willing they would be to update their category on a 7-point Likert scale extending from "very willing" to "not willing at all" (see Appendix). In Experiment 1, participants responded to the following type of prompt: "*Scientists announce they have made a new discovery –*

Table 1
Category rating scales.

Boundedness	Some category boundaries are absolute. It is clear whether something is a member of one category or the other. Some category boundaries are graded. Membership varies by degrees. Please rate the degree to which the following distinction is absolute or graded.
Objective	Some category boundaries are subjective or conventional; they depend on culture, background, or personal opinion. Others are objective; they depend on facts and are independent of culture, background, or personal opinions. Please rate the degree to which the following scientific distinction is conventional or objective.
Discoverable	Some scientific categories are decided by stipulation or legislation: scientists decide what members are in one category and which are in the other. Categorization may be decided by vote or consensus. Other scientific categories are decided by discovery or data: scientists discover what members are in one category and which are in the other. Please rate the degree to which the following scientific distinction is stipulated or discovered**.**
Meaningful	Some categories are meaningful. Knowing something is a member of one category and not another tells you a lot about it, i.e., one can make a lot of predictions or learn a lot on the basis of the category. Please rate the degree to which the following distinction is meaningful or not.

Note. These are the exact prompts the participants read for the four category structure ratings in all experiments. These measures were chosen to tap into essentialist-like beliefs about a category.

they have discovered a previously unknown astronomical body. Scientists announce that this new astronomical body belongs to one of the following categories and not the other. How willing are you to update your category to include the new member?" For each trial (essentialist and deference measures) the category pair was highlighted afterward using real-world examples: for example, "*Planet (e.g., Earth, Saturn, Mars) versus dwarf planet (e.g., Pluto, Ceres, Eris).*" These examples were included to help clarify what updating implied (updating one's list of exemplars in the category). Overall, we focused on willingness to update because it highlights what is at stake in deference; in any case of deference, participants must assimilate expert category judgments into their own category judgments. Thus, this prompt highlighted that updating categories occurs whether experts classify new discoveries or revise old judgments. It also makes clear that deference occurs to the extent that participants are willing to assimilate this new information.

Results and discussion

Participants rated all category boundaries on four measures of essentialist beliefs: objectivity, boundedness, meaningfulness, and discoverability. As all of the measures tapped into how much any given category boundary was construed as demarcating natural kinds, we created a composite measure. A principal component analysis on the four measures supported this reduction strategy: eigenvalues and a scree plot suggested a single component, loading moderately on all measures (0.35–0.60), which explained 39% of the variance. This supports a reduction of the four measures into a "natural kind-iness" or essentialism approximation. Note, however, that although all of these measures significantly correlated with each other ($p < 0.001$), the correlations tended to be small to moderate (0.17–0.46), reflecting that each measure tapped into a different component of natural kind representation. Finally, consistent with our selection of stimuli from presumed natural kinds, all measures were significantly above the scale midpoint-4.0 (collapsing across items): objectivity ($M = 4.97$, $SD = 1.36$), $t(57) = 5.41$, $p < 0.001$, $d = 0.71$, boundedness ($M = 4.92$, $SD = 0.88$), $t(57) = 7.98$, $p < 0.001$, $d = 1.05$, meaningfulness ($M = 5.41$, $SD = 0.97$), $t(57) = 11.09$, $p < 0.001$, $d = 1.46$, and discoverability ($M = 4.59$, $SD = 1.27$), $t(57) = 3.56$, $p < 0.001$, $d = 0.47$.

dependent variable and the composite essentialism measure as the predictor (fixed effect), and participant and item as random effects. P-values were derived from Satterthwaite's approximation via the *lmerTEST* package in R (see Kuznetsova, Brockhoff, & Bojesen Christensen, 2015). As expected, essentialism predicted higher levels of deference, $b = 0.10$, $SE = 0.03$, $p = 0.006$, $= 0.08$.

Some past research suggests that deference is more fragile for participants with less education (Proctor & Keil, 2006). Therefore, we tested whether our results generalized across education level. In support of the generalizability of these results, there was no overall effect of education level on deference, $b = 0.18$, $SE = 0.19$, $p = 0.342$, $= 0.09$, and no significant interaction, $b = 0.07$, $SE = 0.04$, $p = 0.099$.

These results validate that the current methods are appropriate for testing our hypotheses. First, they demonstrate that there is enough variability among the items to detect a relationship between essentialism and deference. Also, our results are consistent with past predictions (Gelman, 2003) and results (Diesendruck & Gelman, 1999; Malt, 1990) that deference should track essentialist intuitions. This suggests our design is consistent with past methods of tapping into deference.

Experiment 2

In Experiment 1, we found that participants defer to experts when experts classify an object for the first time (Diesendruck & Gelman, 1999; Malt, 1990). We also found that higher essentialism predicted greater deference. In Experiment 2, we tested whether there would be a systematic difference in how people defer in cases of novel classification and in cases of re-classification or revision. We hypothesized that deference would be radically different in the context of revision. Because category revision conflicts with intuitive theories about natural kinds – namely, that their boundaries are objective and clear-cut – we expected people to tend to reject revision. By reject, we mean that they would be less likely to defer and more likely to be unsure of whether the past or current judgment is correct.

The major change from Experiment 1 was the replacement of the novel discovery deference prompt with three new prompts about category revision. This produced three separate conditions

We predicted that deference would be high in the context of novel discovery for essentialist categories. Because we sampled from presumed natural kinds/high essentialist categories, we predicted the average levels to be high, which was confirmed (Fig. 1): Participants indicated levels of deference substantially higher than the midpoint of the scale ($M = 5.93$, $SD = 1.09$), $t(57) = 13.74$, $p < 0.001$, $d = 1.77$. Past accounts of deference predict that deference should be highest for categories or category boundaries construed in essentialist terms (Gelman, 2003). To test this proposal, we analyzed the relationship between essentialist intuitions and deference using a multi-level model with deference as the

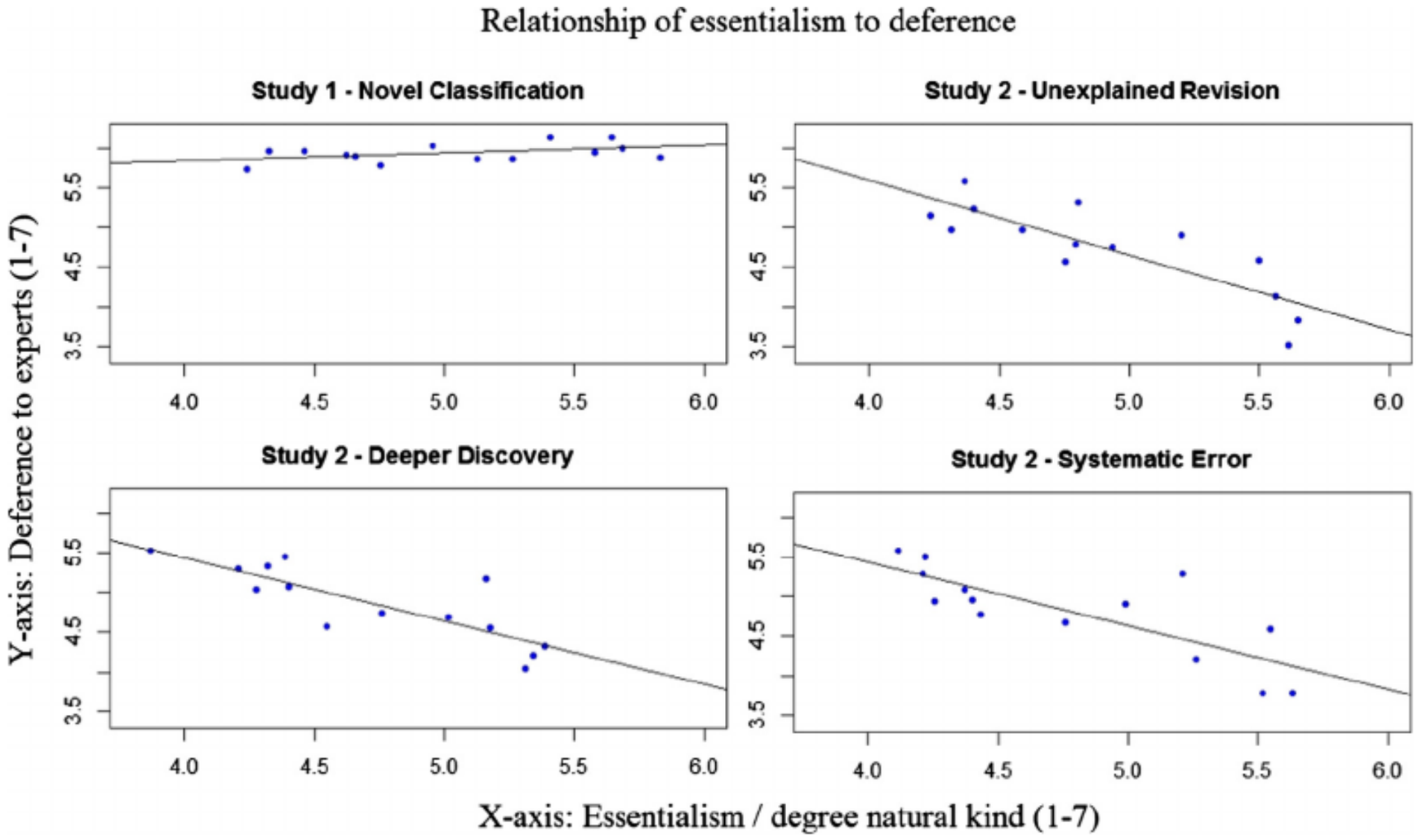

Fig. 1. Relationship between essentialism (degree natural kind) and deference when individual item averages are plotted. Reported results are based on a multi-level model treating item and participant as random effects.

(*Systematic Error*) presented a case in which scientists realized they had previously made large-scale errors in their understanding of a subject. The two explained-revision cases both present typical means by which scientists update their understanding of the reality of the world.

We hypothesized that all three types of category revision would contradict participants essentialist intuitive theories; thus, we expected no condition differences. Essentialist intuitions vary as a function of the stimuli, which were matched across conditions. These types of revision all contradict essentialist intuitions because experts had access to data diagnostic of the category when they made their initial judgment. These types of category revision are normal parts of science, so we hypothesized that participants would only reject them to the degree they conflicted with essentialist intuitions (measured at the item level) rather than for alternative concerns (that may have varied between conditions).

Participants

Overall, 169 adults participated on Amazon Mechanical Turk (who had not participated in Experiment 1). We aimed for approximately 50 participants per condition in keeping with Experiment 1. 59 adults were in Condition 1 (*Unexplained Revision*), 59 in Condition 2 (*Deeper Discovery*), and 49 adults in Condition 3 (*Systematic Error*). The median level of education was some college. 37%

(between-subject), each including a different backstory as to why revision occurred. We manipulated the backstory to ensure that our results were not contingent on any one sub-type of expert revision. If people reject revision because of their intuitive theories, then they should reject any type of expert revision that conflicts with essentialist intuitions. In one condition (*Unexplained Revision*), participants were asked whether they would update their category after an expert re-categorized an entity without an explanation. In a second condition (*Deeper Discovery*) explored a case where scientists improved their understanding of a subject and realized there were deeper/subtler ways of classifying objects. A final condition

covery: "*Scientists have recently discovered a deeper, previously unknown property that separates lions and tigers. In making this discovery, scientists tell you at least one of the animals you believe to be a member of one category was actually a member of the other category. How willing are you to update your category?*" The second was *Systematic Error,* where revision was justified through error correction: "*Scientists have recently discovered there was an error in categorizing a specific member of one of the following categories. There had been a systematic error in the data; upon re-examining the data scientists announce that the animal in question was actually a member of the other category all along. How willing are you to update your category?*"

Directly after the prompt, the categories in question were specified. For the tiger-lion block, for example, participants were provided the following example: Tiger (e.g., Siberian, Bengal, and Sumatran) *versus* Lion (e.g., West African, Barbary, and Asiatic).

Results and discussion

We predicted that revision would conflict with essentialist intuitions. Therefore, rather than readily deferring to experts (as participants did in Experiment 1), we predicted a reduction in average levels of deference – especially because the stimuli were construed in essentialist terms on average. To test this, we performed a cross-experiment comparison comparing levels of deference in Experi-

of participants fell into this category; additionally, 34% had bachelor degrees or equivalent, 17% had only high school degrees, and 10% had graduate or professional degrees.

Design and procedure

There was three conditions that varied category revision backstory. Otherwise, the design and procedure was identical to Experiment 1. In the *Unexplained Revision* case deference was not explained: (for example) "*Scientists tell you that one of the animals you believed to be a member of one category was actually a member of the other category. How willing would you be to update your category?*" There was also two cases of explained revision: The first was *Deeper Discovery* where revision was justified through a novel dis-

condition comparison *ps* > 0.4), reflecting that this relationship generalized across common cases of category revision.

As with Experiment 1, there was no effect of education level on deference, $b = 0.04$, $SE = 0.09$, $p = 0.693$, $= 0.02$, and no interaction between education level and essentialism, $b = 0.005$, $SE = 0.03$, $p = 0.86$. Therefore, unlike past results (Proctor & Keil, 2006), deference was not particularly fragile for participants with lower levels of education.

These results demonstrate that in common real-world cases of category revision people withdraw deference, and that their withdrawal of deference is systematically related to their beliefs about the underlying structure of any given category boundary. Because this pattern reverses the conventional relationship between essentialism and deference, past accounts of deference cannot explain these results. The same beliefs that lead people to strongly defer in Experiment 1 lead people to equivocate in Experiment 2. This happens *even though* all of these categories are completely open to revision in the eyes of science. This reversal is consistent with our hypothesis that people may reject category revision because it is at odds with their beliefs about the nature of categories and expert judgments – namely, that the underlying nature of a category should be causally transparent to a knowledgeable expert and that there is little room for interpretation. Therefore, the more participants believe a category boundary has an essentialist structure (i.e., objective, clear boundaries, meaningful, and discovered), the more they are skeptical of revision. In reality, even categories as seemingly straightforward as water belie an enormous underlying complexity that can make classification extremely difficult (Leslie, 2013).

We also found that the rejection of revision generalized to two of the most common cases of scientific revision, including deeper discovery and correcting systematic errors. Category boundaries that are the most essentialized are assumed to either have been figured out (scientists have determined which entities belong to which categories, what the causal underpinnings of the categories are, etc.) or not. Both deeper discovery and systematic error correction imply the existence of two conclusive category judgments across two time points. Essences are discovered once and for all when an expert has the knowledge, skill, and tools to look underneath superficial appearances and tap into the underlying reality of things. When all this scientific prowess is present, the expert's job becomes a matter of detecting the essence like a litmus test. Revisions contradict this belief; thus people simply do not know which of these conclusions to trust and withdraw deference.

Across two experiments, we find that novel classification and category revision had different relationships to essentialist intuitions. Essentialist intuitions encourage deference when experts newly classify entities, but discourage deference when experts revise previously made judgments. Because past accounts of deference to levels of deference in the current experiment. As predicted, participants were significantly less likely to defer after category revision ($M = 4.83$, $SD = 1.13$) than after novel classification ($M = 5.93$, $SD = 1.09$), $t(102.24) = 6.58$, $p < 0.001$, $d = 0.99$.

We expected that participants are suspicious of revision because it conflicts with their essentialist intuitions. Therefore, we predicted that participants would withdraw deference depending on how essentialist they construe any given category boundaries. To test this hypothesis we used a multi-level model to explore the relationship between deference and essentialist intuitions (model specification identical to Experiment 1). As hypothesized, essentialist intuitions predicted lower deference, $b = 0.21$, $SE = 0.03$, $p < 0.001$, $= 0.15$. Indeed, there was no effect of condition, neither the main effect nor the two-way interaction (all

fered in some of their content, the deference ratings were not strictly comparable across the experiments. To address this concern we employed a within-subjects design and randomly assigned participants to conditions. Second, we decided to replicate our results with a different way of probing deference. The previous prompt, "*How willing would you be to update your category*," makes clear that deference requires updating something about one's previous category judgments (including a new member in the category or switching a member from one category to another); however, the idea of willingness could possibly lead participants to consider the pragmatic value of communicating with others in addition to their actual beliefs about category membership. Specifically, if updating one's category attribution includes updating one's category label, communication concerns are potentially introduced. In Experiment 3, participants were instead asked how likely it was that they would conclude that the object is a member of the category (novel category or revised category). This question more directly taps into the critical component of deference – the effect of expert judgments on people's extension of natural terms and their attribution of category membership to entities in the world. Third, although we did not detect order differences in the above experiments, we greatly reduced the complexity of the experiment by having participants make only eight judgments rather than seventy-five. Together these changes help ensure the results are replicable across different methods and when correcting for the potential limitations of Experiments 1 and 2.

Method

Participants

Ninety-three adults participated from Amazon Mechanical Turk; we doubled the sample size to compensate for the reduction in questions (participants randomly responded to half of the available questions).

Stimuli and procedure

The aim of the experiment was to examine the interaction between essentialist intuitions and different types of expert judgments (independent variables) on levels of deference (dependent variable). Thus, we presented categories low and high in essentialist construal across two types of expert judgments: novel classification and category revision. Participants indicated how the expert judgments would influence their category judgments in each case.

We took the four category boundaries that were consistently rated lowest on the essentialism measures (and thus had low natural kind properties, *Low-NK*) and the four rated highest (*High-NK*). Low-NK stimuli were planet-dwarf planet, dog-wolf, turtle-

ence do not distinguish between these types of expert judgments, deference to experts may be more nuanced than originally described. More critically, these rejections of revisions offer a mechanism by which people's natural concepts can become inaccurate. This equivocation could lead to conceptual inertia as scientific advancements that defy pre-existing knowledge or intuitive beliefs are not readily transferred to the broader public; in turn, people are more likely to persist in outdated conceptual schemes. Given the importance of these results, we wanted to ensure their reliability.

Experiment 3

Three limitations to Experiment 2 motivated a follow-up and replication. First, because participants were not randomly assigned to Experiments 1 and 2 and because the deference prompts dif-

tortoise, and meteor-meteoroid. High-NK stimuli were planet-star, solid-liquid, animal-plant, and reptile-bird. Participants responded only to these eight category boundaries. To further simplify the task, participants did not rate stimuli on category structure (which were not expected to change across experiments); participants only responded to deference prompts.

There are sixteen possible category judgments (eight category boundaries by two types of expert category judgments). Participants were randomly assigned to make eight total deference judgments. For the category revision prompt, participants read the following: "*Scientists announce that an astronomical body you believed was a Planet was actually a Star. How likely are you to conclude that the astronomical body is a Star?*" For the novel classification prompt for the same stimuli pair participants read: "*Scientists announce they've discovered a new astronomical body. They announce the body is a Star. How likely are you to conclude the astronomical body is a Star?*"

We added directions at the beginning of the experiment to clarify the two possible item types. Subjects were told that they would read various hypothetical scientific announcements, and that each announcement fell into two types. They were told that sometimes scientists were categorizing an object for the first time and that, other times, scientists were reconsidering which category an object belonged to. We clarified this to ensure that participants understood that scientists were not *only* revising participants' categories but their own.

Results and discussion

As before, we expected that deference would be relatively high when experts newly classify an entity and relatively low when they revise previous category judgments. To test this first prediction, we used a multi-level model on levels of deference with condition (type of expert judgment) as a fixed effect and participant and item as random effects. Replicating Experiments 1 and 2, deference was substantially higher in the case of novel discovery ($M = 5.76$, $SD = 1.34$) than in the case of category revision ($M = 5.02$, $SD = 1.25$), $b = 0.73$, $SE = 0.13$, $p < 0.001$.

Next, we examined how deference varied as a function of how much participants construed the category boundaries as essentialized. Essentialist intuitions should lead participants to defer when experts newly classify an entity (because experts can tap into its underlying essence) but should lead participants to withdraw deference when experts revise past judgments (because it conflicts with the intuition that experts had already accessed the essence previously). To test this prediction, we used a multi-level model with Condition (Novel Classification vs. Category Revision) and essentialism (Low-NK vs. High-NK) as predictors and participant as a random effect. There were no main effects of condition, $p = 0.341$, or essentialism, $p = 0.128$; there was, however, the predicted two-way interaction between condition and essentialism, $b = 1.06$, $SE = 0.26$, $p < 0.001$, = 0.33. In the category revision condition (replicating Experiment 2), people were significantly less likely to defer for High-NK category boundaries ($M = 4.62$, $SD = 1.80$) than Low-NK category boundaries ($M = 5.37$, $SD = 1.38$), $b = 0.75$, $SE = 0.24$, $p = 0.002$.

In the case of novel discovery, however, deference was higher (though not significantly) for High-NK ($M = 5.84$, $SD = 1.51$) category boundaries than Low-NK ($M = 5.57$, $SD = 1.40$) category boundaries, $b =$ 0.27, $SE = 0.22$, $p = 0.228$. (Recall that in Experiment 1 the relationship of essentialism to deference was on the order of 0.1–0.2 units per scale point, whereas in Experiment 2

classification. The intuitive theory hypothesis was supported by the role of essentialism across Experiments 1–3; namely, that people reject deference in the case of revision in proportion to their essentialist intuitions. Nevertheless, we sought to identify a context of category revision that was compatible with intuitive theories in order to strengthen this conclusion.

Our account suggests that revision is intelligible if experts did not have previous access to the kind of data necessary to diagnose an entity's essence and thus to determine its category membership. One context where category revision can occur without challenging this assumption is one where (1) scientists had restricted access to an object, and (2) data were actively falsified. Therefore, scientists made an incorrect judgment because they correctly and conclusively detected an essence for the wrong object. This is a compelling context because it still requires revision and it still suggests the scientific community can be radically wrong (and indeed can be duped by frauds), but does not conflict with basic assumptions about essences and expert knowledge. We explore this scenario in Experiment 4.

Experiment 4

To test whether people will accept category revision in a context that does not conflict with their intuitive beliefs, we tested deference in the context of fraud. In this case, experts based their category judgment on an alternative specimen (because they were purposefully deceived). Later scientists realized the error and corrected it. In this case, scientists' revision is not based on re-interpretation of the underlying reality of the category member (which they had not actually analyzed before) but on the uncovering of the fraud. We expected participants to defer in this case because even though the scientists had previously been wrong (and indeed, revealed their susceptibility to fraud), their revision does not conflict with essentialist beliefs and thus should not make participants skeptical. We compared the case of fraud to the systematic error condition from Experiment 2 because it was the most structurally aligned (scientists were correcting a previous error in both cases). The only difference was whether experts' original category judgments were based on access to the same underlying structure or whether their revision was based on realizing they categorized the wrong target.

Method

Participants

the relationship was 1–2 units lower per scale point – thus, it is not surprising to find no difference between High-NK and Low-NK category boundaries in the novel classification condition). These results further support the hypothesis that people reject category revision because of their essentialist intuitions; they also support the broader proposal that this rejection is a mechanism for introducing errors into people's concepts. These results are robust across different methods.

So far, our results are potentially consistent with an alternative hypothesis: Perhaps people are opposed to category revision *per se* (rather than because category revision often conflicts with their intuitive theories about category membership and expert judgments). For example, past work demonstrates that people are reluctant to abandon old conceptual schemes even in the face of compelling new evidence (Waldmann & Hagmayer, 2006). Perhaps the mere possibility of revision suggests anything an expert says is subject to change (and thus not worth considering). On the other hand, if people reject category revision because of specific intuitive theories (like essentialism), then participants should not reject revision indiscriminately; they should only reject revision when it conflicts with these intuitive theories about kinds and expert

One hundred and twenty-four adults participated on Amazon Mechanical Turk, divided between two conditions.

Stimuli and procedure

The design of the experiment was to compare deference across two different cases of expert category revision (*Systematic Error* vs. *Fraud*) that we expected to be differently influenced by essentialist intuitions (High vs. Low).

We had the same eight category boundaries as in Experiment 3. There were two between-subjects conditions: systematic error (replicating Experiment 2, Condition 3) and fraud. In the *Systematic Error* condition participants answered the same deference prompt encountered in Experiment 2, Condition 3. In the *Fraud* condition participants answered deference prompts that resembled the following example: "*There is a distant astronomical body that only one member of a single observatory was able to collect data on. Unfortunately, that member was a rogue non-scientist who intentionally falsified the data - sending data from a nearby planet instead. Therefore, the scientific community concluded the object was a planet. Recently however, the fraud was detected and real data from astronomers concluded that instead the object was actually a star all along. How willing are you to update your category (to reflect that the object is a star and not a planet)?*" The details of the deference prompts were specific to each of the eight category boundaries and were varied to be appropriate for the category boundary in question.

Results and discussion

We suspected that fraud would not conflict with essentialist intuitions (because experts had no previous access to category essences), whereas systematic error would conflict (because experts did have previous access to category essences). As predicted, deference was substantially higher in the *Fraud* condition ($M = 5.91$, $SD = 1.29$) than in the *Systematic Error* condition ($M = 5.04$, $SD = 1.12$), $b = 0.87$, $SE = 0.22$, $p < 0.001$.

If fraud truly does not conflict with essentialist intuitions, deference should be as high here as when experts made novel discoveries. To test this, we performed a cross-experiment comparison comparing the *Novel Classification* condition from Experiment 3 with the *Fraud* condition from the current experiment. As predicted, participants did not withdraw deference in the case of *Fraud*; expert deference was as high in the case of correcting fraud as in the case of a novel category judgment, $t(125.02) = 0.724$, $p = 0.47$. This supports our hypothesis that participants only reject revision because it tends to conflict with their intuitive beliefs (and not because they reject revision *per se*).

Next, we again predicted that essentialist intuitions should lead to higher deference when experts' judgments are consistent with essentialist intuitions (as they should in the case of fraud) but lead to lower deference when experts' judgments are inconsistent with essentialist intuitions (as they should in the case of systematic error). To test this prediction we used a multi-level model with Condition (Fraud vs. Systematic Error) and essentialist intuitions (Low-NK vs. High-NK) as predictors and participant as a random effect. There was a main effect of essentialist intuitions, $p < 0.001$, and a main effect of condition, $p = 0.029$, which were further qualified by the expected two-way interaction between essentialist intuitions and condition, $b = 0.64$, $SE = 0.25$, $p = 0.011$. In the *Systematic Error* condition, there was a large effect of essentialism, such that deference was substantially lower for High-NK ($M = 4.70$, $SD = 1.66$) category boundaries than Low-NK ($M = 5.38$,

General discussion

Across four experiments we found support for our hypothesis that people withdraw deference in cases of category revision because it conflicts with their intuitive theories. Revision conflicts with the belief that experts can discern objective and clear-cut categories. In Experiments 1 and 2, we found that participants reliably treated novel classification and category revision in systematically different ways. Novel classification resembled past work on deference – participants readily deferred to experts and were most likely to defer when the categories were the most objective and clear-cut. When experts revised their past category judgments, however, participants withdrew their deference. Often they were equivocal – neither accepting nor rejecting the category judgment. This equivocation was found across different types of common scientific revision and across different methods.

Participants' equivocation in the revision case cannot be explained in terms of past accounts of deference (e.g., Gelman, 2003). Those accounts did not consider revision and presumed that essentialism unequivocally leads to higher deference. In contrast, the finding that participants withdraw deference when experts revise categories conforms to the revised account offered here – participants' essentialist beliefs are double-edged: The belief that natural kinds have real and correct category judgments is why participants seek expert insight at all. However, this same intuitive belief leads participants to dispute the most common cases of category revision. The supposed objectivity and certainty with which natural kinds can be categorized render revision suspicious because experts already confidently asserted a category judgment. To have done so implies that they must have already tapped into the objective underlying structure and determined its correct category membership. Category revision seems to directly challenge the objectivity and certainty of the experts' judgments. Therefore, like a litmus test that returns "acid" at time one and "base" at time two, experts' revising supposedly objective category judgments undermines laypeople's belief that experts are accurate sources of category judgments. In turn, laypeople are unsure of whether old or new category judgments are true (in the same way that one would not know whether the liquid was an acid or a base after testing it with the inconsistent litmus test). These findings provide a fuller picture of deference than offered by past accounts.

This double-edged nature of essentialism and deference pro-

SD = 1.14) category boundaries, *b* = 0.68, *SE* = 0.26, *p* = 0.01, replicating Experiment 2. In the *Fraud* condition, however, there was no difference between High-NK (*M* = 5.93, *SD* = 1.32) and Low-NK (*M* = 5.89, *SD* = 1.38) category boundaries, *b* = 0.04, *SE* = 0.25, *p* = 0.881.

Overall, these results demonstrate that people are not merely opposed to category revision *per se*. Rather, they reject category revision because it generally conflicts with their beliefs about the underlying nature of category boundaries. Fraud does not conflict with essentialist intuitions, so participants deferred to experts duped by fraud as much they did to experts who had made novel discoveries. This pattern of findings supports the proposed mechanism: Highly essentialized category boundaries are real, objective, and clear-cut. Scientists can conclusively discern the category membership of an entity (and an expert category judgment implies this entity was *decisively* categorized).

In the normal case, category revision challenges these beliefs because experts had access to the data diagnostic of an entity's underlying causal structure in both cases. From the perspective of essentialism and other oversimplified causal beliefs, the job of the expert is construed as detecting essences. This underestimates the true causal complexity involved in categorizing entities in the world.

more work will be needed to address skepticism about category revision and to develop more helpful methods of communicating scientific information to the public.

Deference is an important cognitive mechanism. Intuitive theories are a critical component of people's concepts, yet they are often sparse and incomplete (Keil, 2010). Deference is supposed to help fill gaps in sparse theories by lifting the burden of possessing comprehensive scientific knowledge. Accordingly, laypeople can outsource to experts judgments that exceed the capacity of their intuitive theories. Past accounts of deference have been rather optimistic: People's intuitive theories encourage them to defer to expert judgments and this strengthens the scientific accuracy of the entire linguistic community. Unfortunately, we find that these same intuitive theories constrain deference, because simplified causal beliefs suggest that expert judgments are not up for revision.

Deference is predicated on a sociolinguistic division of labor (Putnam, 1973). This stratification may itself be the source of participants' contradictory behavior. In most cases, the division of labor is stratified such that laypeople are not privy to expert judgments directly. Because expert judgments filter downwards through intermediate chains of professionals and arrive preclassified and labeled, this stratification helps isolate laypeople from the messiness of science. This stratification may help to ensure that laypeople defer more than they might if they were more aware of revision. But as much as it may be protective, stratification may contribute to simplified causal beliefs and scientific understanding. First, this stratification might contribute to large lag times between shifts in the scientific community and shifts in the broader public. Second, and more critically, this stratification between experts and non-experts mystifies and caricatures the scientific process, obscuring the true complexity of scientific classification. In turn, this obscuring of the scientific process may explain why participants construe experts and essences in simplistic terms. When laypeople are confronted with how subjective and provisional expert judgments are in practice, they withdraw their deference to experts. Earlier and more consistent exposure to the messiness and intrinsic complexity of the scientific process might

vides a mechanism for understanding a central tension in people's categories; it explains how people can simultaneously behave in externalist networks (attempting to maintain categories that reflect the real world) and how their categories so consistently lack external validity. The findings of the current study add to an increasing understanding of how people's intuitive theories constrain conceptual change and scientific accuracy (Lane & Harris, 2014; Gelman & Rhodes, 2012; Shtulman & Schulz, 2008; Shtulman & Valcarcel, 2012). Prior studies also document a similar tension: The same cognitive mechanisms that promote the accumulation of scientific knowledge can be the greatest barrier to scientific accuracy. Folk biology constrains understanding evolution (Gelman & Rhodes, 2012; Shtulman & Schulz, 2008) and deference to trusted experts can lead to belief in the supernatural (Lane & Harris, 2014). Here we demonstrate a way in which essentialism undermines expert deference in a scientifically disadvantageous way.

Ultimately, the causal underpinnings of most natural categories are highly complex and messy (Leslie, 2013). If participants withdraw deference in the face of this messiness, they may be undermining their very basis for relying on experts. This withdrawal is the true cost of the effect described here. To better improve scientific accuracy and conceptual understanding among laypeople, more work will be needed to address skepticism about category revision and to develop more helpful methods of communicating scientific information to the public.

Deference is an important cognitive mechanism. Intuitive theories are a critical component of people's concepts, yet they are often sparse and incomplete (Keil, 2010). Deference is supposed to help fill gaps in sparse theories by lifting the burden of possessing comprehensive scientific knowledge. Accordingly, laypeople can outsource to experts judgments that exceed the capacity of their intuitive theories. Past accounts of deference have been rather optimistic: People's intuitive theories encourage them to defer to expert judgments and this strengthens the scientific accuracy of the entire linguistic community. Unfortunately, we find that these same intuitive theories constrain deference, because simplified causal beliefs suggest that expert judgments are not up for revision.

Deference is predicated on a sociolinguistic division of labor (Putnam, 1973). This stratification may itself be the source of participants' contradictory behavior. In most cases, the division of labor is stratified such that laypeople are not privy to expert judgments directly. Because expert judgments filter downwards through intermediate chains of professionals and arrive preclassified and labeled, this stratification helps isolate laypeople from the messiness of science. This stratification may help to ensure that laypeople defer more than they might if they were more aware of revision. But as much as it may be protective, stratification may contribute to simplified causal beliefs and scientific understanding. First, this stratification might contribute to large lag times between shifts in the scientific community and shifts in the broader public. Second, and more critically, this stratification between experts and non-experts mystifies and caricatures the scientific process, obscuring the true complexity of scientific classification. In turn, this obscuring of the scientific process may explain why participants construe experts and essences in simplistic terms. When laypeople are confronted with how subjective and provisional expert judgments are in practice, they withdraw their deference to experts. Earlier and more consistent exposure to the messiness and intrinsic complexity of the scientific process might

were bears despite their unusual diet and appearance, and that whales were mammals rather than fish. These are cases of deep conceptual change because scientists update their working theories of the world rather than merely their category extensions. For deference to help laypeople sustain the correct extension of their concepts without acquiring full understanding, experts need to guide people to change the extension of their concepts without dramatically changing their intuitive theories. Providing expertise to laypeople would improve their scientific accuracy, but the function of deference is to sustain scientific accuracy without making non-experts into experts. In other words, deference should help people identify the extensions of their concepts despite their theories being skeletal and underspecified (Keil, 2003, 2010). The work presented here suggests that people's intuitive theories sometimes limit their ability to update the extension of their concepts. Even the shallowest form of conceptual change (updating category extensions) may often require deeper forms of conceptual change (updating intuitive theories). In turn, the extensions of laypeople's concepts tend to lag behind scientists, leading to conceptual inertia in the broader public and the greater chance for scientific inaccuracy.

Appendix A

A.1. Deference prompts

Experiment 1. "Scientists announce they have made a new discovery – they have discovered a previous unknown [*superordinate category, i.e., "animal" for tiger-lion*]. Scientists announce that this new [*superordinate category]* belongs to one of the following categories and not the other. How willing are you to update your category to include the new member?"

Experiment 2. *Unexplained revision:* Scientists tell you that one of the [superordinate category] you believe to be a member of one category was actually a member of the other category. How willing would you be to update your category?

Deeper discovery: Scientists have recently discovered a deeper,

help laypeople eventually overcome these beliefs and be more tolerant of scientific revision. Testing this proposal is a particularly important direction for future work both theoretically and practically. If true, it would provide a mechanism by which to improve laypeople's reliance on experts.

Our findings are not the first to suggest that people's deference may be less principled than one would expect. Braisby (2001, 2004) also finds that people's deference is somewhat limited and that people sometimes rely on their own intuitions. These earlier findings are consistent with our own; people's own intuitions sometimes limit their deference. Indeed, our results help to address an important concern raised by Braisby. Previous accounts of deference argue that essentialist intuitions lead people to defer to expert judgments. Because this link was assumed to be straightforward (and not contextually dependent), the presence or absence of expert deference was interpreted as diagnostic of the presence or absence of essentialist intuitions. Thus, Braisby used participants' lack of deference to argue *against* essentialism. We believe, however, that people withdraw deference *because* they are essentialists. The problem of previous accounts was not in suggesting that people are essentialists, but in assuming that essentialism always entailed expert deference. We demonstrate that essentialist intuitions actually limit deference when scientific judgments undergo conceptual change.

Conceptual change occurs often in the sciences when new knowledge leads to the re-alignment of category boundaries. This type of re-alignment occurred when scientists realized that pandas

previously unknown property that separates [*category one*] and [*category two*]. In making this discovery, scientists tell you that at least one of the [superordinate category] you believe to be a member of one category was actually a member of the other category. How willing would you be to update your category?"

Systematic error: Scientists have recently discovered there was an error in categorizing a specific member of one of the following categories. There had been a systematic error in the data; upon re-examining the data scientists announce that the [*superordinate. category*] in question was actually a member of the other category all along. How willing are you to update your category?

Experiment 3. *Condition 1.* Same Prompt as Experiment 2-*Systematic Error*. *Condition 2.* (Vignettes varied slightly depending on category, this is an example): There is a distant astronomical body that only one member of a single observatory was able to collect data on. Unfortunately, that member was a rogue non-scientist who intentionally falsified the data - sending data from a nearby planet instead. Therefore, the scientific community concluded the object was a planet. Recently however, the fraud was detected and real data from astronomers concluded that instead the object was actually a star all along. How willing are you to update your category (to reflect that the object is a star and not a planet)?

Experiment 4. *Condition 1:* "Scientists announce that a [superordinate category] you believed was a [category-one] was actually a [category-two]. How likely are you to conclude that the [superordinate category] is a [category-two]".

Condition 2: "Scientists announce they've discovered a new [superordinate category]. They announce the [object] is a [category]. How likely are you to conclude the [superordinate category] is a [Category]?"

B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://osf.io/8tqgh/.

References

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

Bloom, P. (1998). Theories of artifact categorization. *Cognition, 66*, 87–93. http://dx.doi.org/10.1016/S0010-0277(98)00003-1.

Bloom, P., & Weisberg, D. S. (2007). Childhood origins of adult resistance to science. *Science, 316*, 996–997. http://dx.doi.org/10.1126/science.1133398.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). Cambridge, MA: MIT Press.

Braisby, N. (2004). Deference and essentialism in the categorization of chemical kinds. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th annual cognitive science society* (pp. 174–179). Mahwah, NJ: Lawrence Erlbaum Associates.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition, 59*, 247–274. http://dx.doi.org/10.1016/0010-0277(95)00698-2.

Braisby, N. (2001). Deference in categorisation: Evidence for essentialism? In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 122–127). Mahwah, NJ: Lawrence Erlbaum Associates.

Coley, J. D., & Luhmann, C. (2000). Domain specific relations between typicality and absolute category membership. Unpublished manuscript. Boston, MA: Department of Psychology, Northeastern University.

Kripke, S. A. (1972). Naming and necessity. In D. Davidson & G. Harman (Eds.), *Semantics of natural language* (pp. 253–355). Dordrecht, Netherlands: D. Reidel Publishing Company.

Lane, J. D., & Harris, P. L. (2014). Confronting, representing, and believing counterintuitive concepts navigating the natural and the supernatural. *Perspectives on Psychological Science, 9*, 144–160. http://dx.doi.org/10.1177/1745691613518078.

Kuznetsova, A., Brockhoff, P. B., & Bojesen Christensen, R. H. (2016). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). Retrieved from <http://cran.r-project.org/package lmerTest>.

Leslie, S. J. (2013). Essence and natural kinds: When science meets preschooler intuition. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (pp. 108–165). Oxford, United Kingdom: Oxford University Press.

Malt, B. C. (1990). Features and beliefs in the mental representation of categories. *Journal of Memory and Language, 29*, 289–315. http://dx.doi.org/10.1016/0749-596X(90)90002-H.

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology, 27*, 41–70. http://dx.doi.org/10.1006/cogp.1994.1011.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language, 31*, 195–217. http://dx.doi.org/10.1016/0749-596X(92)90011-L.

Margolis, E. (1998). How to acquire a concept. *Mind & Language, 13*, 347–369. http://dx.doi.org/10.1111/1468-0017.00081.

Marsh, J. K., & Rothman, N. B. (2013). The ambivalence of expert categorizers. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 984–989). Austin, TX: Cognitive Science Society.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–193). Cambridge, MA: Cambridge University Press.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242–279. http://dx.doi.org/10.1016/0010-0285(87)90012-0.

Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences, 21*, 55–65. http://dx.doi.org/10.1017/s0140525x98000405.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316. http://dx.doi.org/10.1037/0033-295X.92.3.289.

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language, 54*, 407–424. http://dx.doi.org/10.1016/j.jml.2005.10.005.

Proctor, C., & Keil, F. C. (2006). Differences in deference: Essences and insights. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive*

Diesendruck, G., & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin & Review, 6*, 338–346. http://dx.doi.org/10.3758/BF03212339.

Dupré, J. (1981). Natural kinds and biological taxa. *The Philosophical Review, 90*, 66–90. http://dx.doi.org/10.2307/2184373.

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

Gelman, S. A., & Rhodes, M. (2012). "Two-thousand years of stasis": How psychological essentialism impedes evolutionary understanding. In K. S. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 3–21). New York, NY: Oxford University Press.

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory & Cognition, 35*, 1785–1800. http://dx.doi.org/10.3758/BF03193510.

Jylkkä, J. W. (2008). Theories of natural kind term reference and empirical psychology. *Philosophical Studies, 139*, 153–169. http://dx.doi.org/10.1007/s11098-007-9107-y.

Jylkkä, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology, 22*, 37–60. http://dx.doi.org/10.1080/09515080802703687.

Kalish, C. W. (2002). Essentialist to some degree: Beliefs about the structure of natural kind categories. *Memory & Cognition, 30*, 340–352. http://dx.doi.org/10.3758/BF03194935.

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

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7*, 368–373. http://dx.doi.org/10.1016/S1364-6613(03)00158-X.

Keil, F. C. (2010). The feasibility of folk science. *Cognitive Science, 34*, 826–862. http://dx.doi.org/10.1111/j.1551-6709.2010.01108.x.

science (pp. 1974–1979). Mahwah, NJ: Lawrence Erlbaum Associates.

Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy, 70*, 699–711. http://dx.doi.org/10.2307/2025079.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, MA: Cambridge University Press .

Schwartz, S. P. (1978). Putnam on artifacts. *The Philosophical Review, 87*, 566–574. http://dx.doi.org/10.2307/2184460.

Schwartz, S. P. (1979). Natural kind terms. *Cognition, 7*, 301–315. http://dx.doi.org/10.1016/0010-0277(79)90003-9.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology, 52*, 170–194. http://dx.doi.org/10.1016/j.cogpsych.2005.10.001.

Shtulman, A. (2013). Epistemic similarities between students' scientific and supernatural beliefs. *Journal of Educational Psychology, 105*, 199–212. http://dx.doi.org/10.1037/a0030282.

Shtulman, A., & Harrington, K. (2016). Tensions between science and intuition across the lifespan. *Topics in Cognitive Science, 8*, 118–137. http://dx.doi.org/10.1111/tops.12174.

Shtulman, A., & Schulz, L. (2008). The relation between essentialist beliefs and evolutionary reasoning. *Cognitive Science, 32*, 1049–1062. http://dx.doi.org/10.1080/03640210801897864.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition, 124*, 209–215. http://dx.doi.org/10.1016/j.cognition.2012.04.005.

Sloman, S., & Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes, 18*, 563–582. http://dx.doi.org/10.1080/01690960344000035.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition, 74*, 149–175. http://dx.doi.org/10.1016/S0010-0277(99)00071-2.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology, 53*, 27–58. http://dx.doi.org/10.1016/j.cogpsych.2006.01.001.