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 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, respectively). 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 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; 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