Original Articles

The explanatory effect of a label: Explanations with named categories are more satisfying

Carly Giffin *, Daniel Wilkenfeld, Tania Lombrozo

University of California, Department of Psychology, 3210 Tolman Hall, # 1650, Berkeley, CA 94720, United States

A B S T R A C T

Can opium's tendency to induce sleep be explained by appeal to a "dormitive virtue"? If the label merely references the tendency being explained, the explanation seems vacuous. Yet the presence of a label could signal genuinely explanatory content concerning the (causal) basis for the property being explained. In Experiments 1 and 2, we find that explanations for a person's behavior that appeal to a named tendency or condition are indeed judged to be more satisfying than equivalent explanations that differ only in omitting the name. In Experiment 3, we find support for one proposal concerning what it is about a name that drives a boost in explanatory satisfaction: named categories lead people to draw an inference to the existence of a cause underlying the category, a cause that is responsible for the behavior being explained. Our findings have implications for theories of explanation and point to the central role of causation in explaining behavior.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

> First Doctor: "Most learned bachelor whom I esteem and honor, I would like to ask you the cause and reason why Opium makes one sleep?"
> Bachelor: "The reason is that in opium resides a dormitive virtue, of which it is the nature to stupefy the senses."
> Chorus: "Well, well, well, well has he answered! Worthy, worthy is he to enter into our learned body. Well, well has he answered!"

In this well-known passage from *Le Malade Imaginaire*, Molière invites us to question whether appealing to opium's "dormitive virtue" explains why opium makes a person sleep (Molière, 1673/2012). On the surface, the explanation appears nearly circular: what is it to have a dormitive virtue if not to produce sleepiness when ingested, which is the very property the questioner would like to have explained? A deeper look, however, suggests that the explanation may not be as vacuous as it seems. The medieval scholars whom Molière aimed to mock believed that disposition terms marked particular powers or forces internal to the possessing object (Hutchison, 1991).[1] This example suggests that

an explanation that appears to do little more than furnish a label could actually point to a broader network of beliefs that in fact support genuine explanations.

Across three experiments, we investigate whether explanations that invoke a named tendency or condition are considered more explanatory than those that do not, and, if so, why this is the case. For example, is an explanation for someone's abnormal behavior better if it invokes a name (e.g., "she did X because she has *depathapy*, a tendency to X"), than if it appeals to the tendency directly (e.g., "she did X because she has a tendency to X")? And if so, why is this the case? Does a category label support particular inferences (for instance, concerning some stable, causal basis for the behavior being explained?), and do one or more of these inferences offer some reasonable basis for explanation? Below, we review prior work that motivates why a category name could affect the (perceived) quality of an explanation. We then introduce the three experiments we go on to report.

1.1. Psychological background

Several bodies of empirical work shed light on why an explanation that invokes a named category might be judged more explanatory than its unnamed counterpart. First, work with both children and adults suggests that the provision of a category label can have a significant effect on how people conceptualize the category and its relationship to associated properties. Studies find that children prioritize category labels over appearance when making novel

---

* Corresponding author.
  E-mail addresses: carly.giffin@berkeley.edu (C. Giffin), dawilk@gmail.com (D. Wilkenfeld), lombrozo@berkeley.edu (T. Lombrozo).
  [1] We thank an anonymous referee for drawing our attention to this historical context.

inferences about future behavior (Heyman & Gelman, 2000), and that the use of gender labels for objects increases stereotypically gender-consistent behavior (Zosuls et al., 2009). Gelman and Heyman (1999) found that lexicalization – using a noun label to refer to someone who possesses a certain property – caused children to think of the property as more stable over time and across contexts. For instance, children who were told that a child was a "carrot eater" as opposed to a child who "eats carrots whenever she can" were more likely to believe that the child would eat carrots at a later time, and would do so even if her parents did not encourage her to do so.

Studies with adults reinforce the idea that categorical language can support particularly strong inferences. Yamauchi (2005) found that when a person was described categorically (e.g., "Linda is a feminist") as opposed to descriptively (e.g., "Linda believes in and supports feminism"), participants were more willing to draw inferences concerning other attributes that the person might have. Gelman, Ware, and Kleinberg (2010) found that when category labels were embedded in generic statements (e.g., "Zarpies hate ice cream"), participants represented the category in more "essentialist" terms, as reflected in a battery of subsequent tasks including measures of within-category property generalization and stability. Effects of category labels also extend to classification and memory tasks: adults learn named categories more quickly than unnamed categories in simple category learning experiments (Lupyan, Rakison, & McClelland, 2007), and introducing labels seems to support mental representations that are more categorical (Lupyan, 2012) and prototypical (Lupyan, 2016). Together, these findings suggest that category labels can have a powerful effect on how categories are represented and on the inferences they are taken to support.

One proposal is that the provision of a category label could signal that the category is a *kind*, and that kinds in turn license inferences about underlying causal essences. Ahn, Taylor, Kato, Marsh, and Bloom (2013) put forth this proposal and report evidence supporting its latter component: they found that when a category was introduced as a non-arbitrary classification of people or objects, participants were more likely to agree with the statement that there is something shared by all and only members of the category that *causes* other features of category members. In two of their studies, the provision of a category label was one of the properties used to manipulate whether participants would construe a category as a kind versus an arbitrary category. For example, one of their items was a mental disorder that, in the *kinds* condition, was described with the name "BLV":

> There is a mental disorder called BLV that about 500 people have. The official diagnostic criteria for BLV disorder is to display the following three symptoms: has difficulty remembering new information, requires excessive attention, and always chooses solitary activities.

In the *arbitrary categories* condition, participants instead read:

> There are some people in the world who have difficulty remembering new information. There are others who require excessive attention. And there are others who always choose solitary activities. There are some people who have both the 1st and 2nd symptom, some who have both the 2nd and 3rd symptom, and some who have the 1st and 3rd symptom. And it just so happens that there are about 500 people on Earth who have all three symptoms.

They found that participants in the *kinds* condition reported a significantly greater likelihood that "there is a single cause underlying these three symptoms that all and only [these individuals] have (whether or not we know what that cause is)." While the kind

versus arbitrary category manipulation involved several cues beyond the provision of a category label (such as being a known mental disorder with "official diagnostic criteria"), it's plausible that the category label contributed to the belief that the category had some causal basis. If this is correct, then an explanation that appeals to a named category could be judged better because the category is taken to be a causally-essentialized kind that supports causal explanations.

A second body of work sheds light on how adding additional information, even seemingly-vacuous information, could improve the perceived quality of an explanation. Explanations are not only judged better when they are longer (Weisberg, Taylor, & Hopkins, 2015), but also when they contain scientific jargon. In particular, laypeople find circular explanations for psychological behavior significantly better when the explanations additionally contain neuroscience that experts judge to be superfluous (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008; see also Trout, 2008). For example, one group of participants judged an explanation for the curse of knowledge that stated it "happens because subjects make more mistakes when they have to judge the knowledge of others." A second group judged an explanation that additionally stated that it "happens because of the frontal lobe brain circuitry known to be involved in self-knowledge." Adding this additional information led novices, but not experts, to judge the explanation more satisfying.

The effect of adding potentially superfluous claims to a scientific explanation is not restricted to neuroscience; subsequent work has shown that people have a "reductive" bias, generally favoring explanations that contain reductive scientific content over those that do not, with the largest effects found when the augmented explanations are otherwise poor (Hopkins, Weisberg, & Taylor, 2016). Extending these findings to our research questions, it could be that adding a name serves as a cue to implicit reductive content (e.g., signaling that an explanation for behavior is grounded in neuroscience or biology), and/or that there are relevant experts or authorities that underwrite the explanation.

For the domain of mental disorders, additional evidence supports the idea that construing a category in more scientific or reductive terms could affect the way it is conceptualized (Ahn, Proctor, & Flanagan, 2009; Haslam & Ernst, 2002). For instance, Ahn and colleagues found that even trained clinicians have strong beliefs about the biological versus psychological etiology of mental disorders, with more biological disorders more likely to be treated with medication (Ahn et al., 2009). More "biological" disorders are also more strongly essentialized, supporting stronger beliefs in a common cause underlying all cases (Ahn, Flanagan, Marsh, & Sanislow, 2006). Other work finds that phenomena believed to be genetic are thought to be immutable and homogenous (Dar-Nimrod & Heine, 2011). If adding a name in an explanation for abnormal behavior supports a more biological or essentialized construal of the behavior or its basis, it could be that concomitant reductive or essentialist assumptions support an elevated sense of explanatory satisfaction.

In sum, while prior work has not investigated the role of category labels in explanations, work on each half of this conjunction (that is, on category labels only or on explanation judgments only) supports several hypotheses. It could be that people find explanations that appeal to named categories more satisfying than those that do not because (a) the name implies greater stability in associated attributes across time or individuals, (b) the name supports stronger inferences about other attributes, (c) the name implies the presence of a (causal) essence, (d) the name supports a representation with more categorical boundaries, (e) the name evokes more prototypical instances of the category, (f) the name implies more reductive content, (g)