Forthcoming in Cognitive Science. Penultimate draft, please cite published version.

Against Teleological Essentialism

Eleonore Neufeld, University of Illinois at Urbana-Champaign

Abstract

In two recent papers, Rose and Nichols present evidence in favor of the view that humans represent category essences in terms of a telos, such as honey-making, and not in terms of scientific essences, such as bee DNA. Here, I challenge their interpretation of the evidence, and show that it is directly predicted by the main theory they seek to undermine. I argue that their results can be explained as instances of diagnostic reasoning about scientific essences.

Keywords

Psychological essentialism, telos, teleological essentialism, scientific essentialism, diagnostic reasoning, causally-structured concepts, causal model theory

Acknowledgements

For comments and helpful discussions, I'm very grateful to Rick Cooper, Guillermo Del Pinal, Bob Rehder, and two anonymous reviewers for *Cognitive Science*.

1. Introduction

According to psychological essentialism, we represent a host of categories as having an underlying, unobservable ***essence*** that gives rise to the observable traits and behavior of category members. The most prominent version of psychological essentialism is placeholder essentialism (Gelman, 2004; Medin & Ortony, 1989). According to placeholder essentialism, when we essentialize a category, we do not have to know what exactly the essence in question is—we just have to represent the category as having ***an*** essence, whatever it turns out to be (Gelman, 2004; Medin & Ortony, 1989; Rips, 2011). If we acquire the relevant knowledge—say, in the process of formal education—we can elaborate the 'placeholder' and replace it with what we take to be the actual essence. A common view is that the placeholder gets elaborated in terms of what is, according to ***science***, causally responsible for the outward features of a category. For example, when we gain the relevant scientific knowledge, we replace the essence placeholder of our water concept with H_2O. The reason for this is straightforward: according to psychological essentialism, the essence is the ***cause*** of category members' observable features, and it is often science that tells us what exactly such causes are.

Rose and Nichols (2019, 2020) call this hypothesis about how the essence of a category gets elaborated 'scientific essentialism' (henceforth: SEH). They reject scientific essentialism, and instead defend the hypothesis of ***teleological essentialism*** (henceforth: TEH). According to TEH, humans elaborate essences in terms of an Aristotelian ***telos***: the ***function*** or ***purpose*** of a thing which, by hypothesis, defines a category.[1] Lions ***are for*** going to the zoo, clouds ***are for*** raining, mountains ***exist to*** give animals a place

[1] Rose and Nichols also classify telos as an object's ***ultimate cause***, but insofar as this cause is, simultaneously, its goal, end, or purpose, all of which refer to effects, I take the locution of 'cause' in 'ultimate cause' to be metaphorical.

to climb. Rose and Nichols (henceforth: R&N) present various studies, most of which use variations of classic paradigms from the essentialism literature, that, in their view, strongly support TEH. On the basis of their findings they conclude "that people operate with a teleological view of essences" (2019, p. 14).

My aim is to argue that the data R&N present do *not* favor TEH over SEH. To the contrary, their findings are straightforwardly predicted by SEH. If humans represent categories as structured according to SEH, basic principles of diagnostic reasoning *predict* that they should follow precisely the inference patterns R&N report. Note that my focus here is on their 2019 paper, but the points extend to their 2020 paper.

2. The Data

R&N presented participants with different variations of classic essentialism paradigms such as Keil's transformation paradigm (Keil, 1989) and systematically manipulated the telos (i.e., by hypothesis, the essence) of the object in question. For example, for bees, the manipulated telos was *make honey* or *pollinate flowers*. For spiders, the telos *spin webs* or *catch insects* was changed in the critical conditions.[2] The reasoning is that if telos affects participants' category judgements, we have evidence that participants represent essence in terms of telos (cf. Rose and Nichols, 2019, p. 3). I'll briefly describe the first experiment of R&N's 2019 paper to illustrate their main strategy.

In experiment 1, participants were presented with a vignette describing scientists that perform superficial operations on animals. For example, participants were told that scientists performed an operation in which they removed a bee's antenna, wings,

[2] The telos of objects had previously been determined in a pre-study, in which participants were asked questions such as "What is the true purpose of bees?" and "What is the true purpose of spiders?"

changed the number and length of its legs, and so on. The participants also saw a picture of a bee representing the animal before the operation (fig. 1a), and a picture of a spider-looking animal representing the animal after the transformation (fig. 1b).

Figure 1a. Picture presented to participants in experiment 1 in Rose and Nichols (2019) of the bee before the operation.

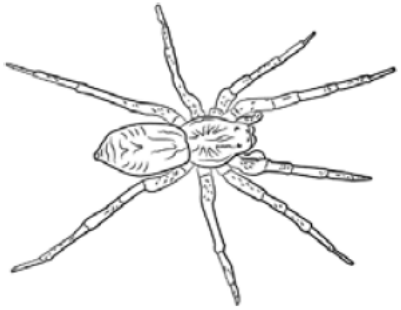

Figure 1b. Picture presented to participants in experiment 1 in Rose and Nichols (2019) of the animal after the operation.

Next, depending on the condition, the participants were given the following information about the telos of the animal:

> ***Telos changed:*** *After running some tests, they found that the thing after the special operation didn't pollinate flowers or make honey. Instead, it only spun webs to catch insects and eat them.*
>
> ***Telos preserved:*** *After running some tests, they found that the thing after the special operation didn't spin webs to catch insects and eat them. Instead, it only pollinated flowers and made honey.*

Thus, one group of participants learned that the telos has changed, and the other group learned that it stayed the same. R&N found that, when the bee was transformed to look like a spider, participants judged that the animal is a spider *when the telos was spinning* (= telos change), but was judged to be a bee *when the thing was still pollinating flowers* (= telos preserved). R&N interpret these results as supporting TEH over SEH.

3. Against Teleological Essentialism

In this section, I challenge R&N's interpretation of the evidence. I first give a brief overview of diagnostic reasoning, and then show that their results are predicted behaviors under SEH in diagnostic reasoning tasks.

3.1 Diagnostic Reasoning

The observable features of an object provide us with important evidence about its unobservable properties. Diagnostic reasoning is our ability to make "inferences from observed effects to (as yet) unobserved causes of these effects" (Meder & Mayrhofer, 2017, p. 433). The ability to infer unobservable properties based on observable ones is indispensable and ubiquitous in everyday categorization tasks.

If a concept is causally structured, the principles we use to make diagnostic inferences about the unobservable properties of an object given its observable ones should be guided by the causal relations represented in the concept. According to SEH, natural kind concepts have a causal structure: the unobservable 'placeholder' or scientific essence causally generates the superficial perceptual properties of natural kind members. Observing effect features should then allow people to diagnostically infer the cause: the underlying scientific essence. And since the scientific essence is usually construed as necessary and sufficient for category membership, humans can infer category membership from the essence.

There is a large body of evidence that we indeed do reason diagnostically in the way predicted by SEH (e.g., Fernbach, Darlow, & Sloman, 2011; Kim & Ahn, 2002; Kim & Keil, 2003; Oppenheimer, Tenenbaum, & Krynski, 2013; Rehder, 2010; Rehder & Kim, 2009; see also Meder & Mayrhofer, 2017; Rehder, 2017 for overviews). For example, using a transformation paradigm, Hampton, Estes, and Simmons (2007) asked subjects

whether, for example, a bird that looks exactly like an insect due to exposure to hazardous chemicals has changed category membership. While more subjects judged the animal to still be a bird, a large minority judged it to have changed category membership to an insect. Importantly, they found that subjects' answers depended on what they inferred about the animal's underlying causal structures. The justifications members of the second group provided made clear that they didn't judge on the basis of appearance. Instead, many used the change in effect features as evidence to ***diagnostically infer*** changes in deeper cause features: "participants reasoned that because both the behavior and the appearance had changed, there must be a new set of causal principles acting within the creature" (Hampton et al., 2007, p. 1790). For example, when subjects reasoned about a giraffe, the fact that not only their appearance, but also their behavior changed made participants infer changes in "deep causal processes within the animal's nervous system" (Hampton et al., 2007, p. 1791). The internal changes, in turn, signaled a change in category membership (to a camel). In contrast, subjects who judged that the animal didn't change its category appealed to the fact that it produced offspring from its original category, and inferred, from this, the absence of internal changes (e.g., to the animal DNA). Hence both groups used observable features to diagnostically infer the state of essentialist causal-scientific structures and determined category membership on that basis.

In fact, the classical studies by Keil (1989) are also an excellent example of diagnostic reasoning based on SEH. SEH predicts that under normal conditions, observable effect features give us diagnostic evidence for their underlying causes. However, when we know that the effect features have been generated by external background causes, and not the causal essence, the inference from effect features to underlying causal es-

sences is defeated. Hence, effects that normally provide evidence for underlying features can be "explained away"– a hallmark of diagnostic reasoning–by external manipulations, which is exactly what Keil found.[3]

Notice a crucial difference between Keil's study and the one by Hampton and colleagues. In Keil's study, subjects were told that the observable features were externally manipulated. In contrast, Hampton's study did not involve an explicit manipulation of observable features: all changes were the product of some sort of *mutation*. Thus, the possibility remained that the changes were due to a change in internal aspects of the category members elicited by the mutation.

3.2 Diagnostic Reasoning and Teleological Essentialism

Let us apply our insights from the last section concerning diagnostic reasoning about scientific essences to R&N's (2019) findings. Throughout the experiments, participants were told that animals undergo some sort of transformation that goes so far as to change animals' normal behavior. *SEH predicts that participants can take this as diagnostic evidence that the scientific essence has, too, undergone a transformation during the surgery.* Since numerous effect features that are normally caused by a bee-essence are not present anymore, participants reason that the underlying scientific essence must have been changed too.

The results in R&N's experiment 1 are straightforwardly predicted by SEH. Recall that participants were presented with a bee that underwent an operation making it look just like a spider. In the critical conditions, the animal either maintained or changed its

[3] See also Oppenheimer, Tenenbaum, & Krynski (2013) for a direct investigation of discounting and augmenting in diagnostic reasoning.

telos. The authors found that the telos manipulation "produced a very large effect on categorization judgements, with participants agreeing that the thing was a spider when its telos was changed […] and with participants agreeing that the thing was a bee when its telos was preserved" (p. 6). Given what we learned about diagnostic reasoning, proponents of SEH shouldn't be surprised by these results. Participants knew that the animal underwent a scientific surgery, but they weren't given any explicit information about whether the 'scientific essence' (e.g., bee DNA) was affected by the special surgery. At the same time, they got information about which observable effect features of the essence changed. SEH predicts that participants make use of the observable information to diagnose the presence or absence of the 'scientific essence'. Since all features that are normally caused by the scientific essence were absent, participants had no diagnostic evidence for the presence of the original scientific essence (e.g., bee DNA). Moreover, in contrast to Keil's classical paradigm, R&N's vignettes did ***not*** include an external causal intervention on tele; rather, the scientists in the story found that the animal ***itself*** generated the relevant telos-behavior "after running some tests". This is crucial because the absence of a manipulated telos as part of the cover story means that ***the telos can be used*** to infer the scientific essence. Thus, SEH predicts that the evidence will be used to infer the corresponding scientific essence.[4]

SEH predicts the outcomes of R&N's other experiments in a way that is structurally analogous. Experiment 2 described how a bee's insides were replaced with a spider's, varied the presence of the telos (i.e., making honey/spinning webs), and found that

[4] This point is especially pertinent given the high diagnosticity of functional features (Lombrozo & Rehder, 2012). Interestingly, Lombrozo and Rehder (2012) carried out many experiments identifying why functional features (akin to telos) have high category diagnosticity. Unfortunately, R&N leave this work undiscussed.

participants' judgements greatly depend on whether the animal has the bee or spider telos. Again, SEH predicts these results: the telos provides important diagnostic evidence for the scientific essence, and should thus have a substantial impact on spider/bee judgements. R&N might claim that by replacing the 'insides' of the bee, they replaced any hypothetical scientific essence. However, according to SEH, the outward behavior simply provides evidence as to whether or not the scientific essence has changed. Thus, the telos might simply be taken as diagnostic evidence for the presence of the relevant scientific essence and that scientific essence does not strictly correspond to 'insides'.

In experiment 3, R&N found that category judgements of the animal are significantly influenced by telos preservation and change in an adoption task. They told participants that a bee was put into a spider cage right after it hatched from its egg, and manipulated whether the telos changed or remained the same. Again, assuming we should use observable evidence in order to make diagnostic inferences about unobservable essences, it is not surprising that manipulating the observable evidence should have a noticeable effect on category judgements. But since other evidence still pointed, diagnostically, in the direction of a bee essence in the "telos changed" condition, the evidence was less clear-cut. Correspondingly, the ratings in the "telos changed" condition were substantially more mixed than for the "telos preserved" condition, with an average rating for the "telos preserved" condition of 1.48, and 4.52 for the "telos changed condition", where 1 = it is definitely a bee, and 7 = it is definitely a spider. Similarly, the multimodal distribution of their data (cf. their figure 3, p. 10) suggests that the subjects were confused about what to do with the conflicting information. One of those modes was at the midpoint of the scale (4.0)—often used by subjects to express uncertainty. The reason for this is clear: the observable evidence favored the

presence of a scientific bee essence in the "telos preserved" condition much more unambiguously, whereas there was conflicting observable evidence in the "telos changed" condition (bee appearance and origin, but spider behavior), accounting for the higher variance in answers and the lower average.

In study 4, R&N found that telos guides category judgements in cross-fertilization scenarios. Just like in experiment 1, participants were first told that a (queen) bee has been superficially altered to look just like a spider. Depending on the condition, the animal was described as either having or not having changed its telos. Subsequently, depending on the condition, the animal's egg was described as being fertilized by ***either*** a spider or a bee. Participants were then asked to rate whether the thing that hatches from the egg will be a spider or a bee. In both conditions—i.e., when the resulting animal's egg was fertilized by a bee and when it was fertilized by a spider—R&N found that the telos of the parent animal after the transformation significantly influenced category predictions about the offspring. Since SEH predicts that we use observable evidence to make diagnostic inferences about scientific essences and category membership, it predicts these results. If observable behaviors serve as evidence of whether the scientific essence has been preserved after the surgical transformation, it will also serve as evidence of what the offspring will be, given that this depends on the presence or absence of the scientific essence.

Finally, R&N's Study 5 combined study 1 with an explicit paradigm. After the category judgements from study 1, participants were asked to rate whether "[the] thing after the changes no longer has the true essence of the original bee". Their aim was to test whether essence judgements generate category judgements. Their mediation analysis showed that telos is a significant predictor for essence, essence is a significant predictor for category, but given the presence of essence, telos is not a significant predictor of category membership. Although there are independent concerns regarding R&N's

methodology in study 5,[5] taking the results at face value, the results of their mediation analysis are, again, completely predicted by SEH. That the relationship between telos and category membership disappears when we control for 'essence' is exactly what SEH would predict. We use telos as diagnostic evidence for the essence, which in turn lets us infer the presence or absence of category membership. Once we know what the essence is, however, knowledge of telos doesn't add anything.[6]

4. Conclusion

The main problem of R&N's study is that they failed to control for the possibility that the observed effects on the category judgements are due to inferences participants draw about a scientific essence. A proper test of the hypothesis that telos constitutes essence should consist in the manipulation of the superficial features only, the hypothesized scientific essence only, and the telos only, in order to assess which of these three manipulations has the biggest effect on categorization judgements, while the setup of the story makes clear that the respective other variables are kept constant. If TEH is correct, it should be far more likely that membership in a category survives change of, say, DNA than change of a given telos. Similarly, if telos remains stable, but the other observable features and the 'scientific essence' change, TEH predicts that category membership shouldn't change either. So according to TEH, a bee that after a transformation has spider DNA and looks like a spider, yet makes honey, should be judged a bee. In contrast,

---

[5] Most theoretical and empirical work on cognitive essentialism assumes that "essence" is a technical term that captures tacit conceptual structures. Thus, the methodology of study 5 seems independently questionable.

[6] Correspondingly, the arrow from telos to essence in figure 6 (Rose & Nichols, 2019, p. 14) is exactly the inferential path based on diagnostic reasoning.

a bee that gets transformed to look like a spider and spin webs, but has bee DNA, should be, contra SEH, judged a spider.

There are a number of broader issues with R&N's proposal. Over many years, a number of cognitive scientists—notably, Woo-kyoung Ahn and her collaborators—have repeatedly found that ***effect features*** are substantially less important for category decisions than cause features (Ahn, 1998; Ahn, Gelman, Amsterlaw, Hohenstein, & Kalish, 2000; Ahn, Kim, Lassaline, & Dennis, 2000; Sloman, Love, & Ahn, 1998; although see Rehder, 2017; Rehder & Kim, 2006 for qualification of the effect). The telos of an object is, as we have seen, the ***goal, end***, or ***purpose*** of a thing.[7] And purposes or goals correspond to certain effects things have: if the purpose of my pen is to write, I expect my pen to have 'writing' as one of its effects. Thus, the pressing question arises how R&N's results relate to the multitude of experiments that report that effect features are ***less*** important for categorization tasks. Why, for example, does Keil (1989) find that preschool children and adults think animals don't change their category when they change functional telos-style features in transformation tasks? Not only does TEH claim that effect features are important for category decisions, they should be the ***most important*** ones, since they constitute the essence of a category. It is not clear how to uphold this hypothesis given the prima facie incompatible results by Ahn and collaborators (among others).

Furthermore, the idea that essences can be thought of as 'scientific essences' has been a common assumption of psychological essentialism for several decades. Hundreds of studies operate under precisely this conception of essentialism—varying

[7] See fn. 1.

genes, insides, atomic elements, and so on. Even if we took R&N's results to unproblematically favor TEH over SEH, going forward, in order for their theory of teleological essentialism to be minimally complete, they would need to explain (or explain away) the vast number of studies that speak in favor of a different hypothesis: that essences are represented in terms of (scientific) causes.

Bibliography

Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, *69*(2), 135–178.

Ahn, W., Gelman, S. A., Amsterlaw, A., Hohenstein, J., & Kalish, C. W. (2000). Causal status effect in children's categorization. *Cognition*, *76*, 35–43.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*(4), 361–416.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in Predictive and Diagnostic Reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–185. https://doi.org/10.1037/a0022100

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409. https://doi.org/10.1016/j.tics.2004.07.001

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory and Cognition*, *35*(7), 1785–1800. https://doi.org/10.3758/BF03193510

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press.

Kim, N. S., & Ahn, W. K. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, *131*(4), 451–476. https://doi.org/10.1037/0096-3445.131.4.451

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory and Cognition*, *31*(1), 155–165. https://doi.org/10.3758/BF03196090

Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive Psychology*, *65*(4), 457–485. https://doi.org/10.1016/j.cogpsych.2012.06.002

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In *The Oxford handbook of causal reasoning.* (pp. 433–457). New York, NY, US: Oxford University Press.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. *Similarity and Analogical Reasoning*, *179*, 179–195. https://doi.org/10.1017/cbo9780511529863.009

Oppenheimer, D. M., Tenenbaum, J. B., & Krynski, T. R. (2013). Categorization as causal explanation: Discounting and augmenting in a Bayesian framework. In *The psychology of learning and motivation, Vol. 58* (pp. 203–231). San Diego, CA, US: Elsevier Academic Press.

Rehder, B. (2010). Essentialism as a Generative Theory of Classification. In *Causal Learning: Psychology, Philosophy, and Computation*. Department of Psychology, New York University: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176803.003.0013

Rehder, B. (2017). Concepts as Causal Models: Categorization. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 347–376). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.013.39

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(4), 659–683. https://doi.org/10.1037/0278-7393.32.4.659

Rehder, B., & Kim, S. W. (2009). Classification as diagnostic reasoning. *Memory and Cognition*. https://doi.org/10.3758/MC.37.6.715

Rips, L. J. (2011). *Lines of Thought: Central Concepts in Cognitive Psychology*. *Lines of Thought: Central Concepts in Cognitive Psychology*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195183054.001.0001

Rose, D., & Nichols, S. (2019). Teleological Essentialism. *Cognitive Science*, *43*(4). https://doi.org/10.1111/cogs.12725

Rose, D., & Nichols, S. (2020). Teleological Essentialism: Generalized. *Cognitive Science*, *44*(3). https://doi.org/10.1111/cogs.12818

Sloman, S., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228. https://doi.org/10.1207/s15516709cog2202_2