Psychological Essentialism and the Structure of Concepts
Eleonore Neufeld, University of Illinois at Urbana-Champaign

Forthcoming in Philosophy Compass. Please cite final version.
This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract

Psychological essentialism is the hypothesis that humans represent some categories as having an underlying essence that unifies members of a category and is causally responsible for their typical attributes and behaviors. Throughout the past several decades, psychological essentialism has emerged as an extremely active area of research in cognitive science. More recently, it has also attracted attention from philosophers, who put the empirical results to use in many different philosophical areas, ranging from philosophy of mind and cognitive science to social philosophy. This article aims to give philosophers who are new to the topic an overview of the key empirical findings surrounding psychological essentialism, and some of the ways the hypothesis and its related findings have been discussed, extended, and applied in philosophical research.

Keywords psychological essentialism, concepts, categories, natural kinds, social kinds, dual character concepts, teleological essentialism, experimental philosophy

1. Introduction: Essentialist Thesis

Imagine a kangaroo that undergoes cosmetic surgery and ends up looking and behaving like a goat. Many would say that this animal, although it looks and behaves like a goat, is nevertheless still a kangaroo. According to psychological essentialism, this is because from a young age, we cognitively represent a host of categories, such as natural kind categories, as having “some hidden, nonobvious, and persistent property or underlying nature shared by members of that kind that causally grounds their common properties and dispositions” (Leslie, 2017, p. 405) and that makes members of the kind “the thing that they are” (Medin, 1989, p. 1476). This doesn’t mean we need to have any concrete views about what exactly the essence is. Standard versions of psychological essentialism assume a ‘placeholder’ view: we often represent categories as having an essence but leave open what precisely the nature of the essence is. Being a member of the kind goat, then, is not a matter
of behaving or looking like a goat, but rather of having some deep, unobservable, intrinsic property: the ‘essence’ of goats.

Following many other theorists in cognitive science and philosophy, we here treat psychological essentialism as a theory about the structure of concepts. Concretely, the structure of essentialist concepts consists of the representation of an essence, a set of weighed typical features that are shared by category members, and a representation of causal relations between the former and the latter (cf. figure 1). This also highlights that the essence in ‘psychological essentialism’ is distinctively causal. The causal essence grounds important modal-inductive inferences associated with psychological essentialism. If an entity $e$ has category essence $k$, which is causally linked to properties $F$ and $G$, you can infer that $e$ will likely have properties $F$ and $G$. However, if you observe that an entity $e$ doesn’t display property $F$, but judge on independent grounds that it has essence $k$, you will judge that it is disposed to have $F$, and would display it if the right conditions were to obtain.

In addition, psychological essentialism usually involves commitments about the way we represent essence and essentialist categories. We treat essences as deep, unobservable, intrinsic and immutable properties that are determined at birth, are causally deterministic and thus have huge

---

1 Since it posits that concepts have internal structure, psychological essentialism is a molecular theory of concepts.

2 In fact, psychological essentialism can be seen as a special case of the Causal Model Theory of concepts (Rehder, 2003b, 2003a, 2010, 2017), according to which concepts represent sets of features and causal relations between them. Psychological essentialism makes specific assumptions about some of the features and causal relations: essentialist concepts represent a common-cause structure, where the common cause is the category essence.
inductive potential. We treat essentialist categories as natural rather than constructed, discovered rather than invented, homogeneous, and as having sharp category boundaries (see table 1 for an overview). Observing that a category has some of these associated properties can then serve as important evidence that the category is represented in an essentialist way. For example, if we have

**Table 1.** Characteristics of essence and essentialist category representations commonly listed in the literature. Some of these markers, especially distinctiveness and sameness, are more controversial as characteristics for essentialism than others. Cf. Gelman (2003), Haslam (2000), Leslie (2013), Rhodes & Mandalaywala (2017), and Rips, (2001, 2011).

- **Natural kinds**
  Essentialized categories reflect real, objective, fundamental distinctions found in nature, instead of conventions that vary across individuals or contexts.

- **Causal power and large inductive potential**
  Category essences are responsible for an object belonging to a category and they causally generate category-typical properties of category members. Thus, category essences have large inductive potential for other category features.

- **Sharp category boundaries**
  Essentialized categories have discrete, not vague or fuzzy, boundaries. A given entity either falls or doesn’t fall in a category.

- **Homogeneity**
  Essentialized categories are homogeneous. Category members will share properties with one another, even if they have other dissimilarities.

- **Immutability**
  Essences and corresponding category membership are immutable and thus stable.

- **Intrinsicality**
  Essences are an intrinsic, not relational or extrinsic, aspect of category members. According to some conceptions of essentialism, essences are also internal.

- **Innate potential**
  (For living entities) Essences are inherited and determined at birth.

- **Hidden and unobservable**
  Category essences are hidden and unobservable. Thus, essences are ‘deep’ and to be distinguished from the superficial, observable features category members exhibit.

- **Distinctness and sameness**
  Different kinds have different essences; same kinds have the same unique essence.
evidence that we represent members of a category as having unobservable, immutable properties, we have some evidence that we have an essentialist representation of the category.\(^3\)

2. Evidence and Paradigms

There’s a large body of evidence that we associate natural kind categories—i.e., categories such as animals (e.g., dog), chemicals (e.g., gold), and plants (e.g., rose)—with the characteristics of essentialism (cf. table 1). In fact, as we see below, we start thinking about these categories in essentialist ways at a very young age. Many theorists have seen this as evidence that psychological essentialism as such is an innate tendency (Atran, 1998; Atran et al., 2001; Gil-White, 2001). Other proposals, in contrast, suggest that essentialism is an effect of multiple other basic, early-emerging cognitive tendencies—e.g., the ability to track objects over time and causal-deterministic biases—which conspire during development and interact with relevant informational input to give rise to full-fledged psychological essentialism (Gelman, 2003; Rhodes & Mandalaywala, 2017; Rhodes & Moty, 2020; Salomon & Cimpian, 2014). As evidence in favor of the latter view, it is pointed out that the different characteristics (or ‘components’) of essentialism seem to have different developmental trajectories. One way or another, all sides to the debate agree that psychological essentialism is a core aspect of human psychology with a fundamental basis, versions of which arise and develop in early childhood and are culturally wide-spread (see §4).

Space constraints prevent us from reviewing the full range of evidence for the presence of psychological essentialism in adulthood and the nuances of its development in childhood. But it will be instructive to get a sense of the evidence by looking at some of the main findings and paradigms that have been used to assess whether a category is associated with the essentialist properties listed above. While the exposition will be roughly guided by the single essentialism components in table 1, it is important to note that the paradigms don’t neatly map to a single one of the essentialism properties, but often display multiple essentialism properties simultaneously, highlighting the interrelatedness of the different components. More detailed reviews and

\(^3\) Importantly, these components of essentialism should be viewed as characteristic features, and do not constitute necessary conditions for a category to be essentialized (cf. Leslie, 2017). Correspondingly, you often find claims of graded essentialism in the literature: a category can be more or less essentialized, depending how many of the components it exhibits.
discussions of the evidence for essentialist category representations as well as the associated developmental trajectories of the different essentialism components can be found in Gelman (2003, 2004), Haslam and Whelan (2008), Mandalaywala (2020), Prentice and Miller (2007), Rhodes and Mandalaywala (2017), Rehder (2017), Rips (2001; 2011), and Rhodes and Moty (2020).

2.1 Triad task: Inductive potential, hiddenness, homogeneity

Markman and Gelman (1986, 1987) used the so-called ‘triad task’ in what became an influential set of experiments. In the task, experimenters create a clash between perceptual resemblance and category membership, and subsequently probe what people choose as the basis for their inductive inferences. In the studies, children were given picture sets with three items per set which belonged to animal or natural substance categories (see figure 2). One of the items closely resembled another item perceptually but was of the same category as the third item. When learning new facts about the items, adults as well as 3- and 4-year-olds would project knowledge from (2b) to (c) with higher likelihood than from (a) to (c), although the item in (a) is perceptually more similar to the item in (c). So, when asked to draw inferences about the leaf-insect, the participants based their inferences on what they knew about the bug, not the perceptually similar leaf. Stimuli included both living and non-living natural kinds. Similar results obtained in control conditions, where Markman and Gelman ruled out that participants simply responded to label similarity. Thus, the results indicate

![Figure 2](image-url) Figure 2. Set of items used in Gelman and Markman’s 1986 and 1987 studies, in which overall similarity in shape and color was pitted against shared category membership. This set comprised (a) a leaf, (b) a beetle and (c) a leaf-insect. When 3- to 4-year-olds heard labels for these items, they were more likely to extend new information on the basis of shared labels than on the basis of overall similarity. Picture source: Gelman (2004)

4 The results have been replicated in various studies (e.g., Davidson & Gelman, 1990; Gelman, 1988; Gelman & Coley, 1990; Gelman, Collman, & Maccoby, 1986; Gelman & O’Reilly, 1988) and have also been extended to younger children, including 1- and 2-year-olds (Graham, Kilbreath, & Welder, 2004; Jaswal & Markman, 2002; Plunkett, Hu, & Cohen, 2008; Switzer & Graham, 2017).
that already young children don’t use superficially visible perceptual information as the basis of their inference. In addition, if the inductive inference isn’t grounded in perceptual similarity, it seems that it must be grounded in a hidden property of the items, which is of higher inductive relevance than perceptual similarity.

2.2 Adoption task and innate potential

The well-known adoption task (also called ‘switched-at-birth task’), introduced in Gelman and Wellman (1991), provides evidence that we treat membership in some categories as fixed at birth, immutable, and as the overriding basis for rich inductions. The strategy of the task is to present participants with a nature/nurture conflict in order to uncover which of them is more decisive for category-related inductions. Participants are presented with a living object that has biological parents of some specified category. At birth, the animal is adopted to be raised by a different family which belongs to a different biological species. For example, an animal whose biological parents are kangaroos is adopted and raised by goats. Next, participants are asked a set of questions about the development of the target animal to find out whether biological parents or the environment are more important for predictions about an animal’s properties. In studies with animals or plants, preschoolers make predictions about an animal’s or plant’s future based on its birthparents, rather than its acquired environment (Gelman & Wellman, 1991). An animal with kangaroo parents that is raised by and among goats and never sees a kangaroo, for example, is judged by children to have a pouch and be good at hopping (instead of climbing) when it grows up. Only very rarely do children base their predictions on environmental factors. Thus, not only does inborn nature ‘trump’ other factors when it comes to predictions about the future, the animal’s nature also doesn’t seem to change as a result of interactions with the environment. Instead, it is immutable and stable.

2.3 Transformation and discovery paradigm: causality, immutability, and hidden structures

Another important source of evidence for essentialism came from Frank Keil’s (1989) famous transformation paradigm. In his experiments, he modified various perceptual properties of an object of a particular category as to look and behave like a member of a different category. For example, in one condition, the subjects were told that a horse was made to completely look and behave like
a zebra through a series of manipulations, such as operation and training by doctors. The relevant task was to decide whether this object was still a horse, or a zebra. The results suggested that, for natural kind categories like animals, kindergartners were more likely to judge according to perceptual appearance and behavior, whereas second graders and adults judged that perceptual appearance didn’t affect category membership. Many have interpreted the results as showing that categorizers attribute some *unobservable constancy* to the animals that isn’t affected by superficial, observable changes (cf. Gelman, 2003).\(^5\) Importantly, the causal structure underlying essentialist concepts (cf. figure 1) also lets us make sense of children’s and adults’ judgements in the transformation paradigms. Superficial features provide evidence for an underlying essence that normally causes the features. But if the superficial features are a result of external intervention, the inference from surface features to corresponding essence is defeated. People and children thus continue to rely on the essence they knew was present before the intervention. This ‘undoing effect’ is a hallmark effect of causal reasoning (Sloman, 2005). In sum, the results of the experiment are evidence that we represent natural categories as having hidden immutable essences that are causally responsible for observable features.

A variation of Keil’s transformation paradigm by Gelman and Wellman (1991, experiment 2) provides evidence that even younger children (4-year-olds) make essentialist judgements in transformation paradigms when the task demands are lower and highlights the special role of *internal* over external properties for category judgements. For a member of a given category, such as a dog, children were asked “What if you take out the stuff inside of the dog, you know, the blood and bones and things like that and got rid of it and all you have left are the outsides?”, or “What if you take off the stuff outside of the dog, you know, the fur and got rid of it and all you have left are the insides?” (p. 225). Children’s identity judgements indicated that insides were more critical for category membership than outsides; they most often judged that an identity change occurred when insides were removed from category members.

---

\(^5\) Keil used more versions of the transformation paradigm, in which different variables of the experiment were manipulated to different degrees (e.g., the changes to the objects were more or less superficial). These affected how early children judged the transformations to not affect category change.
2.4 Category membership vs. typicality: sharp category boundaries and objectivity

Natural kind categories are associated with sharp category boundaries. This has been illustrated in an experiment by Diesendruck and Gelman (1999), in which they found that adults tend to give more absolute category membership ratings for animals than for artifacts. In the experiments, participants were asked to assess to which degree an animal or an artifact is a member of a category and typical of a category. Importantly, for natural kinds, no matter which degree of typicality was assigned to an animal, it was either judged to be a category member or not. In contrast, for artifacts, category membership ratings are not independent of typicality. If an artifact is ‘kind of’ typical, it is also judged to be ‘kind of’ a category member.

3. Conceptual Domains Beyond Natural Kinds

In the last section, we’ve encountered a portion of the evidence that we have essentialist representations of natural kind categories. But much research has been invested in finding out whether we represent categories from other domains essentialistically, with a specific focus on artifacts and social kinds.

3.1 Artifacts

The available evidence about artifact representation has led many theorists to conclude that artifacts are either not represented as essentialized, or essentialized to a degree that is substantially weaker than for natural kinds.6 As described above, adults tend to give more absolute category membership ratings for animals than for artifacts (Diesendruck & Gelman, 1999).7 In addition, in the aforementioned transformation paradigm by Keil (1989), category membership for artifacts seems to run only “as deep as skin”: a pencil that undergoes perceptual change to look and ‘act’

---

7 For more evidence that subjects represent membership in natural kind categories, in contrast to artifact categories, as all-or-none, see Rhodes & Gelman (2009b); Rhodes, Gelman, & Karuza (2014), Estes (2004), and Hampton (2009). These studies specifically tested whether the concepts of personal ignorance and expertise are of greater relevance for natural kind concepts than artifact concepts. The studies suggest that participants think they would be able to categorize given specimen correctly if they had greater knowledge about them for natural kinds, but not artifacts, and, relatedly, that the relevant experts would be able to determine whether a given specimen belongs to a category or not for natural kinds but not for artifacts.
like a hanger is judged by participants to be a hanger. And when explaining properties of artifacts, children often appeal to external rather than internal causes (Gelman, 2003).

An alternative hypothesis is that artifact categories are essentialized, albeit in a different way. Instead of a non-obvious property that causes observable features, we represent artifact essences in terms of an intended function or common effect.\(^8\) The reason why artifacts don’t survive transformation scenarios, then, is because they usually change their function or effect due to the transformation. In support of this view, some research suggests that we privilege functional features when categorizing artifacts. For example, children extend category labels from one artifact to another on the basis of shared function (Kemler Nelson, Russell, Duke, & Jones, 2000). Multiple studies found that subjects regard functions as both necessary and sufficient for artifact categories.\(^9\) Barton and Komatsu (1989) found that functional features were more important than ‘compositional’ attributes (e.g., molecular composition) for the categorization of artifacts, but less important when it comes to natural kind categorization, supporting the view that ‘essence’ has a different functional role in artifact and natural kind concepts.\(^10\) Puebla and Chaigneau (2014) also found that when information about the functional outcome is omitted, physical composition has the highest impact on artifact categorization. This is predicted by the essence-as-function hypothesis: when information about the functional outcome is omitted, physical features provide the best evidence for functional features, as physical features are causally connected to functional effect and thus license a ‘forward-inference’.\(^11\)

Note that accepting the essence-as-function hypothesis for artifact categories would involve a theoretical change of how we think about ‘essence’. In this new sense, a category essence doesn’t have to be the central cause of the typical category properties (cf. figure 1), since ‘artifact essences’ are rather the (intended) functions or effects of artifacts. Instead, what makes something an essence is their unifying role for a category and relative importance for category membership (see §5.2 for further discussion). If the essence-as-function hypothesis was accepted, it would mean that many of the usual essentialism characteristics wouldn’t apply across all essentialism-domains. For


\(^{10}\) Interestingly, and consistent with the essence-as-function hypothesis for artifacts, Wisniewski (1995) found that an artifact object is judged to be a better member of a category when they had feature combinations that would enable the function associated with a category.

\(^{11}\) See Rehder (2017) for discussion.
example, functions usually aren’t intrinsic aspects of entities, and are likewise not hidden or causally responsible for the other properties of an object. Thus, other markers of artifact essentialism would have to be proposed as part of ongoing theoretical developments.

### 3.2 Social Kinds

Research on our representation of social groups has uncovered notable similarities between the ways we represent natural and social kinds. As with natural kinds, we behave as if many social groups are real kinds that have sharp category boundaries, are somewhat “natural”, historically stable and immutable, “real” and not constructed, and allow for rich inductive inferences about physical and behavioral traits of their members. The range of social categories for whichessentialist tendencies have been recorded is wide. Race and ethnicity, gender, caste, sexual orientation, mental illness, and religion, among others, have been associated with psychological essentialism.

One of the first studies that assessed the extent to which we essentialize different social categories was conducted by Nick Haslam, Louis Rothchild, and Donald Ernst (2000). They asked participants to rate to which extent dimensions such as naturalness, stability, discreteness of category boundaries, immutability of category membership, uniformity, exclusivity, and necessity of category features (i.e., dimensions roughly corresponding to the ones in table 1) apply to a category. Among many others, some of the assessed categories were diseases (AIDS patients, cancer patients), dietary

---

14 See Mahalingam (2003).
18 Thus, in contrast to many other paradigms, they explicitly assessed different dimensions along which we essentialize groups. The dimension of discreteness, for example, was described to the participants as follows: “[s]ome categories have sharper boundaries than others. For some, membership is clear-cut, definite, and of an ‘either/or’ variety; people either belong to the category or they do not. For others, membership is more ‘fuzzy’; people belong to the category in varying degrees.”
groups (meat eaters, vegetarians), intelligence groups (smart people, people of average intelligence), races (black, white), religions (Catholics, Jews), and political groups (liberals, Republicans). The results showed that within the 40 social categories that were rated, the category classes of gender, ethnicity, and race as well as the categories Jews, AIDS patients, homosexuals and the tested political groups exemplified essentialized categories.

In the same study, Haslam and colleagues found that essentialism scores were related to social status ratings. In order to assess the social status associated with the social categories, they asked participants to rate the categories “as you think the general public sees it’ … i.e., ‘how valued or favorably regarded are they in our culture” (2000, p. 118). The categories that scored highest on multiple essentialism dimensions received the lowest status ratings. This finding converges with a much-cited hypothesis by Gordon Allport, who conjectured that psychological essentialism is the root of social prejudice (Allport, 1954). “Prejudice”, in this context, refers to the general set of negative attitudes we can have towards social groups, including affective (e.g., a feeling of aversion associated with a group), behavioral (e.g., discriminatory behavior) and cognitive components (e.g., stereotypes). Overall, evidence regarding the relationship between essentialism and general forms of prejudice has been mixed. There has been consistent evidence, however, in favor of a positive relationship between psychological essentialism and a specific form of prejudice: stereotyping (i.e. cognitive forms of prejudice, such the belief that Germans are boring, or an association between being German and being boring). For example, using different questionnaires that assessed participants’ knowledge and endorsements of social stereotypes, Bastian and Haslam (2006) found that essentialist beliefs uniquely predict stereotype endorsement, independent of other predictors of stereotyping. In another study (2007), the authors found that essentialist beliefs predict greater attention allocation to stereotype-consistent, rather than stereotype-inconsistent, information. Indeed, this positive association between essentialism and stereotyping is directly predicted by psychological essentialism, since it entails expectations of strong inductive potential and within-category homogeneity (Rhodes & Mandalaywala, 2017).19

19 See also Haslam and Whelan (2008), Mandalaywala and Rhodes (2017), and Mandalaywala (2020) for a general overview on the relationship between psychological essentialism and different forms of prejudice. See, e.g., Bastian and Haslam (2006, 2007), Levy and Dweck (1999), Levy, Stroessner, and Dweck (1998), Dar-Nimrod and Heine (2006), Coleman and Hong (2007), Pauker, Ambady, and Apfelbaum (2010), Yzerbyt, Corneille, and Estrada (2001), Pauker et al. (2016), and Gallagher and Bodenhausen (2021) for findings that support a positive relationship between essentialism and stereotyping. See Ritchie and Knobe (2020) and Bailey and Knobe (2021) for alternative views. The latter study present evidence that (biological) essentialism is correlated, but doesn’t cause, stereotyping and prejudice.
There’s yet another crucial finding that Haslam and colleagues (2000) presented in their seminal study. Their results indicated that essentialism’s different components don’t co-occur in an all-or-nothing fashion. Instead, the study revealed that the different components cluster into two main dimensions, each of which seems to have different implications for prejudice and stereotyping. The authors described one of those dimensions as the “natural kind” dimension, which “indexes the degree to which a social category is understood to be specieslike” (Haslam et al., 2004, p. 69), and includes the components discreteness, naturalness, stability, and necessity. The other dimension, described as the “entitativity” dimension, indexes how much a social category is “judged to have an underlying ‘real-ness’” (ibid.) and includes the components uniformity, informativeness, inherence, and exclusivity.20 The data also indicated that the relationship between essentialism and social status described above looks more intricate if we pay attention to the two dimensions. The entitativity dimension, by itself, was significantly correlated with social stigma, while the correlation between the natural kind dimension and stigma was substantially less strong. Notably, however, the categories that were especially stigmatized tended to score high on both essentialism dimensions.

Although more evidence for a two-factor structure has since been documented, and there have been some theoretical discussions about the phenomenon (see especially Mandalaywala, 2020 and Haslam et al., 2004), research on this topic has been relatively sparse and predominantly focused on social categories.21 One important question that needs careful examination is how theoretically significant the two-factor structure is. A tempting conclusion is that “psychological essentialism” in fact tracks two different cognitive constructs. It is not clear that this conclusion is warranted. Note that we shouldn’t expect that the different dimensions in table 1 are, in general, highly correlated in the first place. We should only expect that they are correlated conditional on a category being essentialized. Consider a disease, Dis, which causes the symptoms sweaty feet, dry mouth, and low blood pressure. Conditional on someone having Dis, we expect the symptoms to be correlated. But if we look at all possible diseases, we probably won’t find that the symptoms are correlated, because the symptoms can occur in different constellations with other diseases. For

---

20 For an informative overview on the history of the concept of entitativity within psychology, see Haslam et al. (2004).
example, you can have low blood pressure if you have Dis, but also if you have a variety of other diseases that don’t cause you to have either sweaty feet or dry mouth. Similarly, we treat the different essentialism-components as symptoms of psychological essentialism. Conditional on a category being essentialized, though, the dimensions should correlate. Haslam et al. tested a wide variety of categories—they didn’t only include categories that they independently thought to be essentialized. Thus, it is to be expected that some of these categories—precisely the ones that aren’t paradigmatically essentialized—will rank high on one dimension but not the other.

In fact, Haslam and colleagues (2004) themselves favor a view according to which “psychological essentialism” tracks the presence of both dimensions—entitativity and natural-kindness. Support for this position, they claim, stems from the finding that entitativity and natural kind beliefs are dynamically linked, “in that manipulations of beliefs in each set generated inferences to beliefs in the other set” (Haslam et al., 2004, p. 75), and that “the social categories often claimed to be prone to essentialist understandings—racial, ethnic, gender, and sexual minorities—[were] judged to be both natural kinds and entitative, [and] the interaction of these two dimensions strongly predicted evaluative judgements of our sample of categories” (ibid.). Future research and theoretical discussion have to clarify many outstanding questions in this domain, e.g. to which extent the two-factor structure extends to natural kinds, what is the developmental trajectory of the two dimensions, and what is the relationship between the dimensions and behavioral consequences.

4. Context-sensitivity and cross-culturality

Psychological essentialism is context-dependent along multiple dimensions. With respect to social categories, there is cross-cultural variability in whether children essentialize race and ethnicity: while US-children have more essentialist beliefs about race than children from Israel, children in Israel essentialize religious and ethnic categories more strongly than US-children (Diesendruck, Goldfein-Elbaz, Rhodes, Gelman, & Neumark, 2013). Own group membership also affects essentialism. Within US-contexts, racial minority children seem to display more essentialist

---

22 For an overview, see Rhodes and Mandalaywala (2017) and especially Pauker, Tai, and Ansari (2020) and Haslam, Holland, and Karasawa (2014).
judgements at an earlier age than white children. Plausibly, this is because racial minority parents are more likely to discuss and emphasize race and its social significance with their children than white children. Surrounding culture and exposure to diversity also have an influence on the development of essentialist representations. A study that compared children of various ages from a liberal university town to children from a rural town that is socially and politically more conservative as well as ethnically and racially more homogeneous found that essentialist gender beliefs declined with age in participants from the more liberal town, while they stayed stable in the more homogeneous, conservative community (Rhodes & Gelman, 2009). And while younger children from both participant groups didn’t exhibit essentialism about race, older children and adults from the conservative community were more likely to essentialize race. Many other studies found an effect of diversity exposure on essentialist category representations.

Interestingly, while social essentialism, like natural kind essentialism, generally seems to emerge quite early in childhood (between age 3 and 5); there are remarkable developmental differences between essentialism about different social target categories—specifically, gender and race. Essentialism about gender categories seems to emerge earlier and be much more context-insensitive than essentialism about race categories; in fact, essentialist beliefs about gender seem to emerge earlier than essentialist beliefs about any other social categories, and persist throughout development into adulthood. Children view gender boundaries as being objective and marking different kinds of people, and they expect the birth gender and sex to determine behavioral and biological properties later in life—even when growing up in the opposite gender environment. Children furthermore display essentialist beliefs about gender even when older children and adults in their community display flexible or anti-essentialist views, or when the children have transgender identities. In contrast, young children (between 3 and 4 years) have flexible beliefs about race.

26 For extensive overviews, see Pauker, Tai, and Ansari (2020) and Rhodes and Mandalaywala (2017).
27 Cf. Rhodes and Mandalaywala (2017); Pauker, Tai and Ansari (2020) for discussion.
category boundaries, and the view that races form objectively distinct kinds of people doesn’t develop until ages 7–10, indicating that the development of race essentialism is protracted compared to gender essentialism.\textsuperscript{31} Some have interpreted the special development of gender essentialism as evidence that gender concepts, but not other concepts, have unique constraints in cognitive development.\textsuperscript{32}

The finding that social essentialism seems to vary contextually raises the question of how universal psychological essentialism as such is. On the one hand, the literature on psychological essentialism often emphasizes that essentialist cognitive structures have been found in all populations that to date have been studied (see, e.g., Gelman, 2003). These include (but aren’t limited to) populations in North America, West Africa, Hawai‘i, Turkey, Brazil, China, South Africa, Mongolia, Israel, Japan, Australia, Guatemala, South Korea, Kazakhstan, India, and Ireland.\textsuperscript{33} A common conclusion drawn from this is that psychological essentialism is a cross-cultural, if not universal, phenomenon. At the same time, we’ve just seen that psychological essentialism is context-sensitive along multiple dimensions. Importantly, this doesn’t only apply to social, but also to natural categories (Gelman, 2003; Olivola & Machery, 2014). Based on this contextual variation, Olivola and Machery (2014) have argued that psychological essentialism is neither a universal nor a stable human phenomenon; thus, they conclude, there’s no reason to think the human mind is ‘fitted’ to essentialize. However, the fact that the extent to which we essentialize a given category varies along several dimensions doesn’t imply that the phenomenon of psychological essentialism is not universal. The universality-claim is that the tendency to essentialize some categories is found in all cultures, not that all cultures essentialize the same categories. Correspondingly, many theorists who claim that essentialism is an outgrowth of innate tendencies hold that the categories that end up essentialized strongly depend on environmental input.\textsuperscript{34}

\textsuperscript{31} Kinzler and Dautel (2012), Rhodes and Gelman (2009), Roberts and Gelman (2015).
\textsuperscript{32} For an overview of psychological essentialism in relation to different categories, see Rhodes and Mandalaywala (2020) and Pauker, Tai, and Asari (2020).
5. New Theoretical Developments

5.1 Teleological Essentialism

Psychological essentialism, recall, is usually committed to a ‘placeholder’ view, according to which we often don’t have a precise representation of what the essence of a category is. What happens when we acquire relevant knowledge about a category—e.g., in the course of education? According to this view, we may then ‘elaborate’ the placeholder. If we learn that H2O is what makes water potable, transparent, etc., we elaborate the essence correspondingly. That is, we usually elaborate the essence in scientific terms. The reason for this is clear: the essence of a category is the cause of its members’ features, and science is what usually tells us about underlying causal features of natural kinds.

Recently, Rose and Nichols (2019, 2020) have argued in favor of a different view of essence. According to their version of psychological essentialism—teleological essentialism—we don’t elaborate essence in scientific terms, but rather via an Aristotelian telos: i.e., the purpose or function of a thing. For example, if bees are for honey making, the essence of bees is making honey. Rose and Nichols performed various experiments in which they manipulated the telos of an animal (e.g., ‘making honey’ for bees) to test whether it affects category judgements. For example, in one of the experiments, participants were told that scientists performed an operation in which they removed a bee’s antenna, wings, changed the number and length of its legs, and so on. After the surgery, the animal looked like a spider. In the ‘telos preserved’ condition, the animal still made honey, and in the ‘telos changed’ condition, the animal only spun webs post-surgery. They found a large effect for telos on category judgements: participants judged that the animal was a spider when the telos was spinning, and a bee when the telos was honey-making. Rose and Nichols concluded on this basis that humans represent essence in terms of telos (Rose & Nichols, 2019).

In response, Neufeld (2021) has argued that Rose and Nichols’ results are directly predicted by standard psychological essentialism. According to her, their experiments don’t rule out that participants simply use information about the ‘telos’ to reason diagnostically about scientific essences, which is why telos affects category judgements. When we reason diagnostically, we make “inferences from observed effects to (as yet) unobserved causes of these effects” (Meder & Mayrhofer, 2017, p. 433). Thus, it is possible that participants treat the telos as observable effect, and use it to make inferences about its cause—i.e., the scientific essence responsible for the effect.
Recent empirical evidence lends preliminary support to the diagnostic reasoning view (Joo & Yousif, 2021).

5.2 Dual Character Concepts and Generalized Essentialism

Another new area of research is the relationship between psychological essentialism and concepts with value-based components—particularly, dual character concepts. The notion of ‘dual character concepts’ was originally introduced to capture a class of concepts that are special in that they are associated with two separate criteria for classification—one descriptive, and one normative (Knobe, Prasada, & Newman, 2013). A paradigmatic example is scientist. We can imagine that someone satisfies all concrete criteria for being a scientist (analyzing data, doing experiments, working in a lab, etc.) but fails to satisfy an abstract value associated with it (being invested in the quest for truth). In this case, we might say that they are, in a sense, clearly a scientist, while at the same time failing to be a scientist in another sense (e.g., failing to be a ‘true scientist’). Newman and Knobe (2019) argue that these concepts are too in the domain of psychological essentialism and thus form a unified class with paradigmatic examples of psychological essentialism, such as zebra. In the case of value-based dual character concepts, some deeper value or ideal (e.g., ‘quest for truth’) constitutes the essence of the category, which is realized via the concrete, observable characteristics (e.g., doing experiments). Newman and Knobe call this type of psychological essentialism ‘platonic essentialism’. In the case of well-known essentialist categories, such as zebra, an underlying hidden property constitutes the category essence, which causes the concrete, observable properties category members share. Newman and Knobe call this ‘causal essentialism’. Importantly, both types of essentialism are unified, according to them, via the mechanism of general essentialism: “the tendency to try to explain observable features in terms of a further unifying principle” (Newman & Knobe, 2019, p. 2).

Knobe and Newman support this hypothesis—i.e., that the same general essentialist mechanism is responsible for both phenomena—by pointing to a number of psychological similarities between them. One source of evidence they appeal to comes from Tobia, Newman, and Knobe (2020), who found that natural kind concepts such as water are also associated with a

---

35 See Reuter (2019) for an overview. See also Machery and Seppälä (2011) for a similar theory.
36 See also Del Pinal and Reuter (2017).
dual character pattern: i.e., with two separate criteria for categorization. For example, when given Twin-Earth-style vignettes (Putnam, 1975), participants judge that there’s a sense in which the liquid on Twin Earth is water, while simultaneously judging that it isn’t water in another sense. Thus, as with value-based dual character concepts, natural kinds can be categorized according to their superficial features only, and according to their causal essence only. These results aren’t only interesting insofar as they support Newman and Knobe’s ‘general essentialism’, but also because they suggest that paradigmatic essentialized concepts, such as natural kind concepts, are dual character concepts. While more research is needed to put Newman and Knobe’s ‘general essentialism’ to the test, recent results look promising: Bailey, Knobe, and Newman (2020) found that value-based forms of essentialism and ‘causal essentialism’ elicit similar beliefs and inductive inferences.

Note that Newman and Knobe’s ‘general essentialism’ seems to incorporate the theoretical change of the notion of ‘essence’ and essentialism we discussed in the context of artifact categories and the essence-as-function hypothesis (cf. §3.1). According to ‘general essentialism’, an essence doesn’t have to be an underlying cause for observable surface features. Instead, what makes something an essence is their unifying role for a category. Insofar as the characteristics in table 1 are only associated with ‘causal essentialism’, if we accept the framework of general essentialism and other ‘species’ of essentialism (e.g., platonic, causal, social, or artifact essentialism), further work would have to uncover and clarify what the components of other forms of essentialism are.

6. Objections

Psychological essentialism has been subject to both theoretical and empirical criticisms. Influential empirical evidence against psychological essentialism was presented by Barbara Malt (1994). She asked subjects to judge how much H$_2$O they think is in liquids called “water”, and liquids not called “water”. Malt found that estimated amounts of H$_2$O did not monotonically correspond to whether something is labeled “water”, contrary to what seems predicted by psychological essentialism. For example, tea was judged to consist of 91% H$_2$O, and pool water was judged to consist of 81% H$_2$O.

37 A complicating factor is that in previous studies, Knobe, Prasada, and Newman (2013) found that natural kind categories such as animals are not associated with a dual character pattern. Thus, more research is needed to find out what the relationship between natural categories and dual character concepts is.
Malt’s results and their interpretation are controversial among philosophers and psychologists. For example, the evidence does not seem problematic for versions of psychological essentialism in which the presence of the essence is necessary but not sufficient for category membership. Malt’s methodology also seems to assume that we have direct verbal access into our representation of essence, which is independently questionable.

In addition to these empirical problems, Michael Strevens (2000) has offered an important theoretical criticism of psychological essentialism. His basic claim is that there is no need to posit a representation of ‘essence’ as part of our conceptual structure in order to explain our causal-essentialist inferences. His simpler ‘Minimal Hypothesis’ that merely attributes beliefs in the relevant causal laws between, for example, tigers simpliciter and stripes to humans, can explain the data equally well. Recall the ‘triad task’, in which Gelman and Markman (1986) showed children pictures of one gray squirrel, one rabbit, and a ‘kaibab’: a squirrel that looks like a rabbit. After being told that the gray squirrel eats bugs and the rabbit eats grass, the children judged that the kaibab eats bugs rather than grass. The essentialist explanation of children’s behavior is that after seeing a kaibab, children (a) represent it as being a squirrel since they are told it is one, (b) infer that the kaibab has the squirrel essence, and (c) infer, from their belief that the squirrel essence causes bug-eating, that squirrels eat bugs. Strevens points out, however, that the essence is of no importance for the inference to go through—we can account for the data that children believe the kaibab to eat bugs simply by children’s belief “that something about being a squirrel causes an animal to eat bugs” (Strevens, 2000, p. 156). Thus, essences do no explanatory work.

A possible response would be to ask what it means to say that humans simply represent that ‘something about the squirrel’ causes it to eat bugs—without having any opinion of what that something is, or of what makes this causal law true. Arguably, this proposal simply collapses into the placeholder version of psychological essentialism, according to which we represent that something about the squirrel, while typically having no opinion of what that is, causes it to eat bugs. This ‘something’ is precisely what essentialists call the ‘essence’. In a response piece, Ahn and colleagues (2001) furthermore note that we have independent evidence that humans have representational constraints on what this ‘something’ is. For animals, for example, it is something internal and something that is transferred and inherited by birth. Contra Strevens, then, (all)

---

opinions about essence do not “appear relatively late, perhaps largely as a result of formal education” (Strevens, 2000, fn. 14, p. 163).\footnote{See Strevens (2001) for a response. See also Strevens (2019) for recent objections to psychological essentialism.}

7. Psychological Essentialism and Social Philosophy

Apart from issues in the philosophy of mind and cognitive science, the framework of psychological essentialism has been applied and is of importance to a variety of other issues in philosophy, including issues in ethics and moral psychology, epistemology, philosophy of race and gender, as well as metaphysics and personal identity.\footnote{See Appiah (2018), Dar-Nimrod et al. (2011), De Freitas, Cikara, Grossmann, and Schlegel (2017, 2018), De Freitas, Tobia, Newman, and Knobe (2017), Kornblith (1993), Leslie (2013), Mallon (2016), Martin & Heiphetz (2021), Mizrahi (2014), Newman, De Freitas, and Knobe (2015), Rips, Blok, and Newman (2006), Rose (2021), Strohminger, Knobe, and Newman (2017), Strohminger and Nichols (2014).} In this last part of the review, we focus on some of the ways philosophers have drawn from the framework of psychological essentialism to explain many phenomena of interest within social philosophy.

One area of application is the philosophy of dehumanization. According to Livingstone Smith (2011; 2014, 2016), when we dehumanize others, we cognitively attribute a less-than-human, or subhuman, essence to them (and, correspondingly, fail to attribute a human essence to them). In these cases, we make an appearance-reality distinction along the lines of psychological essentialism: although someone might look like a human, they have a subhuman ‘essence’. In addition, ‘subhumans’ rank lower than humans in the moral hierarchy of kinds that, in Livingstone Smith’s view, is an implicit part of our moral-psychological system and determines the moral value of natural kinds.\footnote{Note that Livingstone Smith does not draw on empirical evidence when making this speculation (which he himself acknowledges, and attributes to lack of existence of such literature) (Livingstone Smith, 2016).} This process plays a distinct moral-psychological role, insofar as essentializing a person as subhuman disables inhibitions we’d otherwise have against harming them. Livingstone Smith draws on evidence from the Rwanda genocide, the Holocaust, and other historic atrocities to argue that the phenomenon of subhuman essentialization is systematically present in these contexts. In a recent paper, Neufeld (2020) also argues that there is an intimate connection between many acts of cognitive essentialization and dehumanization. According to Neufeld, when we essentialize someone, we represent them as having a ‘nature’ that lawfully disposes them towards certain behaviors and attributes. Thus, we deny their agency and represent them as inherently
constrained in their capacity for self-determination *qua* member of a social kind. Neufeld uses this insight to shed light on the dehumanizing effects of pornography, which, according to Neufeld, systematically depicts women in essentialized ways.

Psychological essentialism has also been used to shed light on multiple phenomena within social philosophy of language. Sarah-Jane Leslie has argued that the framework helps us understand important properties of the linguistic class of *generics*, such as their connection to stereotyping and prejudice. Many generics about essentialized kinds, such as *Mosquitos carry Malaria*, are readily accepted by hearers despite the extremely low prevalence when the attributed property is *dangerous*. Because we want to avoid danger, even if it comes with small likelihood, we quickly accept the generic and attribute the property to the *kind*: it’s in the *nature* of mosquitos to carry Malaria, even if the property isn’t realized at a given moment. Although this kind of mechanism might be innocent in the case of mosquitos, it becomes highly pernicious when applied to social groups. Also here, when a generic attributes a dangerous property to a social group, accepting it commits us to the belief that the *essence* of the group predisposes their members to have the property in question—even if the propensity is extremely low. Among other things, according to Leslie, this explains why many stereotypes and prejudices are extremely resistant to counterevidence.

According to Neufeld (2019), *slurs* also rely on erroneous essentialist mini-theories. They express that members of certain groups have some underlying, causally powerful ‘essence’ which is causally responsible for negative stereotypes associated with a social group. This account doesn’t only predict slurs’ linguistic behavior correctly; it also gives a straightforward explanation for why they dehumanize. According to the mini-theory encoded in a slur, it is just in the ‘nature’ of members of some social group to exhibit certain negative properties. These mini-theories then deny the target’s capacities for autonomy and self-determination. The essentialist account of slurs also makes sense of the otherwise curious fact that slurs occur predominantly as nouns. Research in cognitive psychology and psycholinguistics shows that nouns are particularly prone to essentialize. If the semantics of slurs is essentialist, nouns should be the primary linguistic vehicle through which we communicate their content. In recent work, Ritchie (2021) also makes use of the

---


insight that nouns have been found to be prone to essentialize to argue that this puts important constraints on the prospects for anti-essentialist ameliorative projects—i.e., projects that advocate for conceptual revisions for political or social justice aims. Given that nouns are particularly prone to elicit essentialist views of social groups, Ritchie argues that the continued use of nouns might undermine ameliorative projects with anti-essentialist aims.

8. Conclusion

This paper provided a systematic overview of the empirical and theoretical status of psychological essentialism. Essentialist theories of concepts have gained much traction in psychological and philosophical work on concepts. It speaks to the fruitfulness of the research program that it proposes answers to foundational questions in cognitive science and informs multiple issues in philosophy, and simultaneously leaves open a variety of other puzzles for philosophers and cognitive scientists to solve. These include, but aren’t limited to:

- How do we represent essence? Do we need to posit a representation of ‘essence’ to accommodate our categorization and induction behavior?
- Are there different ‘species’ of essentialism? What is the underlying relational structure of each? What does this mean for our theoretical notion of ‘essence’?
- How do the different components of essentialism develop and cluster? What does this mean for the theoretical notion of ‘psychological essentialism’?
- Are there other category domains we represent as essentialized?
- Does psychological essentialism ‘debunk’ certain philosophical intuitions, e.g., metaphysical essentialism about natural kinds?
- How does psychological essentialism interact with other cognitive capacities, such as perception and language?
- What is the content and structure of positive self-essentialism?
- How does psychological essentialism interact with social ideologies?
- How can we ameliorate problematic essentialist representations about social groups?
- How is psychological essentialism about kinds related to psychological essentialism about individuals.
References


45. https://doi.org/10.1016/j.paid.2017.06.029


