This is a preprint of the following chapter: Martínez-Manrique, F. (2025). Psychological Essentialism and Natural Kinds. In M.J García-Encinas & F. Martínez-Manrique (eds.) *Special Objects: Social, Fictional, Modal, and Non-Existent* (pp. 107-130). Cham: Springer Nature Switzerland, reproduced with permission of Springer Nature Switzerland. The final authenticated version is available online at: https://link.springer.com/chapter/10.1007/978-3-031-82221-6_6

Psychological Essentialism and Natural Kinds

Fernando Martínez-Manrique

Abstract According to psychological essentialism, people divide the world into categories that are seen as possessing deep, underlying properties that account for what is common in members of the category. I examine two ways in which this phenomenon has been used either to debunk or to vindicate essentialism about natural kinds. I argue that neither way affects the essentialist thesis, since they depend on other types of evidence that independently reject/support the thesis. I contend that research on psychological essentialism may play a more direct role in a different argument, which addresses the reality of certain natural kinds. To this end, I will revise the issue of the mind-independence of natural kinds through the concept of unification principle. This concept can offer a criterion of natural kindness that allows certain sorts of mental dependence as constitutive of an objective mind-independent kind. I will apply this idea to the case of race, examining some ways in which findings about psychological essentialism could either debunk or vindicate the existence of such a natural kind, and extend it other putative kinds.

Keywords Psychological essentialism · Natural kind · Unification principle · Mind-dependence · Debunking/vindicating arguments · Race

1. Introduction: the relevance of cognitive science for metaphysics

People hold different sorts of metaphysical beliefs. They may think that everything is material, or they may believe that there are immaterial entities of some kind. They may have different ideas about fundamental metaphysical concepts, such as time, cause, identity, or free will. Some

of those ideas, when properly scrutinized, may turn up to be contradictory[1], but this does not typically affect people's lives, or makes people change their ideas when the contradiction is spelled out –unless they are professional philosophers, and perhaps not even in this case. Studying this 'folk metaphysics' is not the province of philosophy, but –as it is the case of other folk beliefs– a task for psychological research (or perhaps its close associate, experimental philosophy, see Livengood and Machery, 2007, for an articulation of this project). However, knowing something about it may be relevant for philosophical projects, at least in two important respects. First, it could be the case that metaphysical stances of the sorts that are formulated by philosophers are ultimately based on metaphysical intuitions that are closely related to folk-metaphysical intuitions held by their non-philosopher neighbors. If this is the case, studying the latter could be a good way to elucidate the nature of the former. Second, even if the connection between metaphysical and folk-metaphysical stances is not so straightforward, knowing more about the cognitive processes that lead people to make metaphysically-loaded judgments about their world can help to illuminate the resources and limits that metaphysical thinking has at its disposition.

Cognitive science research, then, can be relevant for metaphysics. The most typical way in which this relevance can be put to work is known as the "debunking" strategy (Goldman, 2015). The general idea is to use scientific findings and theories to undermine the basis for a certain philosophical view. In cognitive science, the typical background is a reliabilist approach to epistemically instrumental mechanisms. Some of our cognitive processes seem to be less reliable than others, in the sense of including biases and distortions that mislead people towards the wrong judgments. In this account, to debunk a metaphysical view is to show that the relevant intuitions or beliefs that back the view result from an unreliable process. However, parallel to this mostly negative role, cognitive science could also play a more positive one. If showing that there is an unreliable process behind the formation of certain metaphysical beliefs can cast doubt on them, then showing that all the processes involved are typically reliable may give some support to the metaphysical view in question. I will call this positive role the "vindicating" strategy[2]. Of course, both strategies, particularly the latter, must be handled with caution. For many cognitive processes, it is usually the case that they are reliable when applied to a certain domain yet unreliable when applied to a different one, and it may be very difficult to ascertain

[1] See Schwitzgebel (2014) for a defense of the idea that common sense leads to incoherent metaphysical systems.
[2] Frugé (2019) uses the neologism "unbunking" to refer to this positive role, but I think that "vindicating" is a more descriptive term.

whether a class of metaphysical judgments "belong", even in a loose sense of this word, to that domain or not.

Bringing evidence from cognitive science to debunk a metaphysical view is thus not a straightforward matter: there is no direct route from scientific claims to metaphysical claims. As Paul contends: "Whether your particular observations and your metaphysical theory (…) as a whole are threatened will depend on how such competing claims on evidence are adjudicated, whether you have other sources of evidence, on your particular metaphysical views, and on a holistic assessment of the merits and demerits of competing metaphysical theories" (2016, p. 427). Similar considerations can be made with respect to vindicating strategies: however well supported by empirical evidence a scientific theory is, making a metaphysical claim out of it requires additional philosophical work.

The aim of this chapter is to examine the prospects of applying the debunking/vindicating strategies by means of an extensively researched psychological phenomenon, known as psychological essentialism (Medin and Ortony, 1989; Gelman, 2003). Psychological essentialism is the idea that people divide the world into categories that are seen as possessing deep, underlying properties –an essence– that account for what is common in members of the category. Those properties are typically regarded as the causes of the shared observable features and behaviors of the members of the category. Psychological essentialism appears associated with two related metaphysical notions: one is essentialism tout court, i.e., the philosophical doctrine whose most representative supporters are Kripke and Putnam; the other is the notion of natural kinds, which mark distinctions between categories of objects that exist "out there" in a way that is independent of our minds –i.e., to put it in the customary informal way, that "carve the world at its joints". The study of psychological essentialism could thus help to illuminate the basis for our metaphysical commitments regarding these notions.

The structure of the paper is the following one: First, I offer a summary of the main findings in psychological research that allegedly support the thesis of psychological essentialism. Then I examine two ways in which psychological essentialism has been used either to debunk or to vindicate essentialism about natural kinds. I will spell out the respective arguments to show that neither of them really has good prospects to affect the essentialist thesis, since they depend on other types of evidence that independently reject/support the thesis. I will suggest research on psychological essentialism may play a more direct role in a different argument which addresses the reality of certain natural kinds. To this end, I will revise the issue of the mind-independence of natural kinds through the concept of unification principle (Tahko, 2015). This concept can offer a criterion of natural kindness that allows for some special kinds

of objects: they are objects that belong to objective mind-independent kinds that are partially constituted by certain sorts of mental dependence. I will argue that psychological essentialism may provide such a sort of mental dependence for some natural kinds, and that understanding what sort of dependence is involved throws light on possible debunking/vindicating strategies for those natural kinds. I will apply this idea to the controversial case of race, arguing that it could be regarded as a natural kind if race categorization processes can be characterized as a narrow common cause for the clustering of properties in that kind, as the unification principle proposes. I will examine some ways in which cognitive science findings could either debunk or vindicate the existence of such a natural kind, and extend it other putative kinds of objects.

2. Psychological essentialism

Decades of research on how humans categorize reality have revealed that people tend to divide the world into categories that are seen as possessing deep, underlying properties that account for what is common in members of the category. Those properties are typically regarded as the causes of the observable features and behaviors of the members of the category, especially those features that are typical and distinctive so as to make one category different from others. For instance, horses can be seen as sharing typical properties –a certain shape, a range of sizes, or a sociable temperament– due to the fact that they share "a deeper, nonobvious reality, that there exists some inherent, internal, immutable substance or quality (the "essence") that causes the characteristics that category members share" (Gelman, 2019, p. 315). This hidden essence can be something different in each case: presumably genes in the case of biological kinds, or an atomic substructure in the case of chemical kinds like water. But people do not need to have a particular theory about it –arguably, psychological essentialism existed before people were acquainted with genes, atoms, or other scientific constructs. They can simply refer to the "innards" of the objects as responsible of their distinctive observable properties. Many scholars subscribe to a "placeholder view" of essentialism (Medin and Ortony, 1989), according to which people may have incomplete knowledge about the essential properties of a category but are disposed to assign some "essence" to it. This "essence" acts as a temporary placeholder until more accurate information becomes available that allows them to revise and reorganize their views about the category.

While people's ideas about these essences may vary, they tend to include the beliefs that they are intrinsic properties of the creatures, not extrinsic or relational ones; that they are transmitted from parents to offspring; that they are unalterable and stable over transformations;

that they demarcate sharp boundaries between categories, and that they have a rich inductive potential, providing the grounds for inferences that extrapolate certain features of individuals to members of their same category (Gelman, 2003). Although these beliefs are most closely associated with natural kinds, especially biological kinds (Atran, 1998), psychological essentialism can be found across many domains of categorization. Of particular importance is social essentialism, i.e., the view that certain social categories, such as race, gender, or ethnicity, involve boundaries between deeply distinct kinds of people (Prentice and Miller, 2007; Rhodes and Moty, 2020). But essentialism can be discerned in many other domains, such as emotions (Lindquist et al., 2013), artifacts (Gelman, 2013), or psychiatric disorders (Berent and Platt, 2021).

Psychological essentialism is nowadays a fertile field of research with many open questions. One is to what extent it can be regarded as a universal phenomenon that appears across different cultures (Neufeld, 2022). On the one hand, the way and depth in which people essentialize is affected by different contextual factors, including cultural ones. On the other hand, a tendency to essentialize certain categories has been observed in all cultures that have been studied, even if not all of them essentialize the same categories. Regarding how essentialism is acquired and developed, there is strong evidence that essentialist tendencies come early in development (Gelman, 2003). They can be observed in young children, even if not at the same time for every category. For instance, biological kinds, but also some social kinds such as gender, seem to be essentialized already in 3-4 year old children, but race essentialist distinctions would appear around 7-8 years. This raises a question about the relation between essentialism for biological and social kinds. Some theories regard the former as more basic and the latter derived from it (Gil-White, 2001). Others warn against the tendency of regarding biological essentialism as more fundamental than social essentialism and argue that both arise from a general abstract structure (Newman and Knobe, 2019).

Underlying this discussion, there is the question about the mechanisms that support psychological essentialism. One point of contention is whether they are domain-general or domain-specific, i.e., whether they are the result of fundamental cognitive capacities that are employed for a variety of purposes, or the result of specialized mechanisms that work for different domains with different principles. On the side of domain-specificity, some authors hypothesize an innate folk-biological module underlying universal essentialist propensities in reasoning about animals and plants (Atran, 1998). Exposure to different cultural and environmental inputs will combine with those propensities to determine the particular expressions of such a system in different groups, showing a way to reconcile cognitive

universality with cultural diversity (Medin and Atran, 2004; Hirschfeld, 1996). On the side of domain-generality, Gelman rejects the idea of domain-specific essentializing systems, arguing instead that essentialist categorization is supported by a variety of general systems that are applied in different domains (Gelman, 2003). Among the different domain-general mechanisms that have been proposed, we have implicit cognitive processes such as the inherence heuristic (Cimpian and Salomon, 2014), cues from generic language (Rhodes et al., 2018), or general sortal object individuation capacities (Rakoczy and Cacchione, 2019).

There is also an ethical side in psychological essentialism insofar as it plays a role in dehumanization of certain groups (Smith, 2014). When applied to social categories, such as racial or ethnic groups, essentialist thinking can lead to the conclusion that these groups possess inherent differences that set them apart from the rest of humanity, reducing them to stereotypes that obliterate their common human qualities and, in the extreme cases, characterizing them as less-than-human. It is important thus to understand how psychological essentialism works to make better policies to deal with the harmful consequences, for instance, of racial categorization (Kelly et al., 2010). So cognitive science is valuable as a companion to ethical projects.

The interest of this paper, however, lies in its utility for metaphysics. Psychological essentialism seems to be well positioned to play a role here too. It deals with belief-producing mechanisms of the sort that underlie intuitions that have to do with metaphysical stances, such as the commitment to essences and to the reality of certain kinds. It provides a starting point to construct arguments that question the reliability of those metaphysical intuitions –the debunking arguments– as well as arguments that lend support to those intuitions –the vindicating arguments. In the following section, I will examine an argument of each type that appeals to psychological essentialism to debunk/vindicate essentialism. I will argue that findings from cognitive science do not play a substantial role in them, as the main source of the debunking/vindicating force comes from other domains. Yet I will frame a different sort of argument in which cognitive science research on psychological essentialism can play a stronger, more direct role.

3. Psychological essentialism and metaphysics: debunking or vindicating?

Psychological essentialism is a psychological theory, not a metaphysical one. More specifically, it can be regarded as a psychological theory about the sources of people's metaphysical views. In other words, when we say that the layperson categorizes and reasons in terms of hidden essences, we are attributing her a folk metaphysical standpoint about the deep structure of the

world. We can wonder now whether there is a relation between this folk metaphysics and metaphysics in general. In particular, it seems reasonable to wonder whether psychological essentialism can be used either to debunk or to vindicate metaphysical essentialism. Actually, one can find both strands in the literature.

3.1 The debunking argument against essentialism

Let us consider debunking first. As I pointed out in the introduction, the typical debunking strategy starts from the unreliability of a certain psychological process to undermine the plausibility of a certain metaphysical view that is allegedly supported by intuitions originated from that psychological process. For psychological essentialism to play a debunking role, the argumentative strategy would be roughly this:

> *The debunking argument against essentialism*
> (i) Metaphysical essentialism relies on intuitions of the sort revealed by psychological essentialism.
> (ii) Research on psychological essentialism shows how people's intuitions about categories are based on biases and distortions.
> (iii) Psychological essentialist intuitions are thus based on unreliable processes.
> hence,
> (iv) Metaphysical essentialism depends on unreliable intuitions, so it is an unreliable view itself.

Varieties of this argumentative schema can be found in Leslie (2013) and Gelman (2019). Leslie is more explicit in linking psychological and metaphysical essentialism, given that her goal is to mount an attack on the essentialist views derived from the seminal work of Putnam and Kripke (Putnam, 1975a; Kripke, 1980)[3]. She begins by presenting a general metaphysical picture about reality called Quintessentialism. According to this picture, nature is divided into real kinds, objectively determined by substance-like entities –the quintessences– that are the causal source of the stable and enduring intrinsic properties that characterize the individuals that belong to a kind. She then reviews evidence from psychological essentialism

[3] Leslie acknowledges that, in the case of Putnam, he moved towards a more moderate position. Her point, however, is not an exegetical one, but an examination of the thesis as it is usually understood that derives from Putnam (1975a).

research –which I summarized in the previous section– to show that "quintessentialist thinking is not a local phenomenon, but rather is a pervasive aspect of human psychology" (Leslie 2013, p. 119). In other words, essentialism would constitute a widely shared folk metaphysical stance.

Leslie endorses premise (i) of the debunking argument when she links the intuitions delivered by such a stance with the intuitions behind Kripke/Putnam essentialism. Her point is that the latter appeals to intuitions that are shared by preschool children, and that are traceable to the deep-seated cognitive outlook of psychological essentialism. She contends that if we did not have such an implicit belief set then we would not have the relevant philosophical intuitions[4]. The intuitive appeal of the Kripke/Putnam essentialism is due to an entrenched cognitive bias. To be fair, Leslie is not explicitly arguing that people's intuitions are straightforwardly false. As she contends, the existence of an early-developing implicit belief set does not establish the falsity of the intuitions based on it –it simply gives reason "to scrutinize the intuitions, and look for independent and converging evidence for the conclusions they urge" (2013, p. 109).

Now, what is the source of this "independent and converging evidence"? The answer is to be found in scientific research in the domains where people conjecture the manifestation of essences. Leslie offers an extensive revision of examples from biology and chemistry that provide mounting evidence against the thesis that members of the kinds share something that could be characterized as an essence. In the biological domain, there is a wealth of cases in which phenotypic variation can be the result of very different microstructures. To put but an example, the widely held belief that having two X chromosomes is an essential feature to be a female does not withstand scrutiny. In the chemical domain, regarding by many as the most favorable for essentialist claims, phenomena such as isotopic variation question the possibility of finding a common microstructural essence of substances (Needham, 2008).

In contrast to Leslie, Gelman is not too concerned with attacking a particular metaphysical view, but she can still be regarded as endorsing premise (i) when she states that psychological essentialist biases and distortions "constitute a deflationary account of essentialism as a theory of how the world is structured" (2019, p. 315). In other words, people's cognitive biases lead them to hold metaphysical stances that are an oversimplification of reality. Gelman summarizes those biases and distortions in three respects: people tend to underestimate

[4] When it comes to the referential intuitions that support the sort of Kripke/Putnam essentialism, it is not clear whether they are so widespread among the folk. In studies testing those intuitions there are findings that suggest that people often do not use words in conformity with essentialism (Braisby et al., 1996), and that many people share intuitions predicted by the causal-historical theory, while many others have descriptivist intuitions, with cultural differences playing some role in the differences (Machery et al., 2004).

the variability within a kind, to overestimate category boundaries, and to assume a causal essence shared by members of a kind. Now, to conclude that people's judgments are distorted implies that we have some standard to measure the degree of distortion. Once again, the question is the source of that standard. Similarly to Leslie, the answer is to be found in science. For instance, we conclude that people underestimate variability because their estimations of variation among individuals of a given folk kind are much lower than the findings revealed by evolutionary research; and we conclude that they overestimate category boundaries because they establish sharp distinctions among categories that research shows to be more fuzzy or fluid.

The conclusion thus is that these debunking arguments derive their strength from the normative force of scientific findings. It is necessary to have some normative ground to conclude that people are wrong in attributing essences to members of categories such as biological or chemical kinds. The source of this normative ground comes typically from research on biology and chemistry, respectively. One does not need to show that psychological essentialist tendencies are biased to debunk the view that there are hidden and distinct essences that underlie biological kinds. Rather, debunking goes in the opposite direction: we conclude that people's essentialist views are biased insofar as science disproves the existence of such essences. Metaphysical essentialism for biology is debunked just by evolutionary biology. Since Darwin, the argument goes, evolutionary biology has moved away from essences and their derived Aristotelian notions, such as species or genus. Psychological essentialism may help to understand why people have those entrenched folk essentialist beliefs, but it does not add much to undermine essentialism itself.

3.2 The vindicating argument for essentialism

Let us turn to a possible vindicating role for psychological essentialism. Such an argument would roughly have the following role:

> *The vindicating argument for essentialism*
> (i) Metaphysical essentialism relies on intuitions of the sort revealed by psychological essentialism.
> (ii*) Research on psychological essentialism shows how people's intuitions about categories deliver a sizeable number of right beliefs.
> (iii*) Psychological essentialist intuitions are thus based on reliable processes.
> hence,

(iv*) Metaphysical essentialism depends on reliable intuitions, so it is a reliable view itself.

Now, one might think that this argument is doomed from the beginning. Given what we have just seen about essentialist beliefs being biased, it seems that premise (ii*) is just simply wrong. However, not everyone agrees with this. For instance, Bloom offers an argumentative strategy that points towards such a vindicating role. Let me quote him at length in a paragraph that nicely captures his position:

> "I don't think the human propensity toward essentialism is actually a mistake (…) biologists reject the Aristotelian notion that species are unchangeable ideal types with no intermediate forms (…). It is this sort of ''essentialism'' that is mistaken. But this is much stronger than Lockean essentialism, under which the superficial features of entities are the result of deeper causal properties. Essentialism in this more general form is simply a belief that reasons exist as to why things fall into certain categories: birds are not merely objects that resemble each other but instead have deeper properties in common. This sort of essentialism is rampant in current biological thought (…). Essentialism is an adaptive way of looking at the world; it is adaptive because it is true". Bloom (2000, pp. 152–153)

Even though Bloom's line of argumentation, like Gelman's or Paul's, also relies on judgments from biology regarding the truth of essentialism –at least for biological kinds– it also provides an independent consideration that supports those judgments. The point here is the adaptive character of psychological essentialism. If essentialist categorizations have contributed to the adaptive success of humans, then they must be the product of reliable processes. The argument bears a resemblance to the "no miracles argument" for scientific realism, which holds that this is "the only philosophy that doesn't make the success of science a miracle" (Putnam, 1975b: 73). Analogously, the evolutionary success of humans would be an extraordinary coincidence unless most of our entrenched beliefs about our surrounding reality were right. Hence, psychological essentialism would provide a way of vindicating the general essentialism about the structure of the world that, according to Bloom, "is rampant in current biological thought".

How convincing is this argument? It depends on which beliefs are those that one deems to be right. If they are the beliefs that the observable features of certain objects are the result of deeper causal properties, then they are probably correct, and Bloom is right in contending that they are shared by biologists too. But if they are the beliefs that these deeper causal properties are common to member objects of a kind, and that they characterize what it is to be a member of such a kind, then biology probably parts ways with the folk. Creatures X and Y can belong

to same folk kind and have the same observable features F, and yet the deeper properties that cause F in X could be different from the deeper properties that cause F in Y. As Gelman observes "concepts of natural kinds do not map neatly or unproblematically onto "true" kinds" (2019, p. 316).

The crucial difference between the debunking argument and the vindicating argument lies in the role of, respectively, (ii) and (ii*). It is worth noting that these premises are not, strictly speaking, contradictory. Yet they reveal a tension between reliabilist and unreliabilist interpretations with respect to categorization processes. On the vindicating account, one may safely assume that categorization processes yield reliable outcomes in many everyday instances. For instance, when discovering an exemplar of a new animal, if I observe that it has a carnivore diet, I will typically assume that the rest of the animals that I regard as belonging to the same kind are carnivores too, and I will be right in the overwhelming majority of cases. This inductive reasoning will be also successful with many other quotidian properties –mating habits, ability to fly, or edibility for humans. On the debunking account, however, one may lay the emphasis on the large discordance between the folk picture and the scientific picture of the world. The folk may have many details right, yet their general worldview –e.g., about how those details are to be explained– is utterly distorted. The dispute cannot be resolved until we are clear about what kind of essentialist beliefs are actually attributable to people. Let me briefly examine this question.

3.3 Is psychological essentialism actually essentialism?

At the bottom of the discussion lies the issue of what sort of intuitions we are attributing someone when we label her as "essentialist". Are they very general intuitions that can be easily make compatible with her, mostly correct, everyday inductive judgments? Then one would be more inclined to accept (ii*). Or are they more articulated metaphysical stances that do not mesh well with the established scientific image? Then the balance would lean towards (ii). It is here where premise (i) appears to be relevant. The general point is that there is some connection between, on one side, people's ordinary metaphysical points of view and, on the other side, metaphysical theses of the sorts discussed in philosophical texts. Yet there are reasons to discard the idea that such a connection could be straightforward. After all, metaphysical theses are elaborated by philosophers with a degree of sophistication beyond the qualifications of the layperson. The situation is somehow parallel to the role of scientific judgments in assessing premise (ii): even if, as humans, scientists are under the effect of the same biases as any

individual, scientific research has a degree of sophistication and collective control that goes beyond individual capacities and that confers it the power to correct people's judgments. This is, as I argued, where the normative force of the debunking argument lies.

Now, philosophical investigation is equally sophisticated and controlled –if not by evidence, at least by the public discussion of arguments. Should not we equally assume that philosophical investigation goes beyond individual abilities and that has some power to overcome individual biases? In other words, when students learn about philosophical points of view –such as the arguments and thought experiments that support essentialism– are not they educating their judgments, so that they can revise their previous intuitions, whatever they are? Even if one rejects this idea, i.e., even if one thinks that philosophical intuitions are inextricably linked to ordinary intuitions, one at least should accept that the former are not a mere reproduction of the latter, but an elaboration of them. In other words, folk-metaphysical statements cannot be directly translated to metaphysical statements –a certain amount of philosophical reflection is needed. This suggests that some caution is needed when one labels folk-metaphysical views with metaphysical labels, such as essentialism.

Indeed, not everyone agrees that people categorize in terms of essences. For instance, according to Strevens people connect kind membership with observable properties by means of causal laws and it is unnecessary to attribute them any belief in essences (Strevens, 2000). In response, Ahn et al. contend that people have richer theories about kinds and about how surface features are brought about by inner causes, in a way that allows to regard these causes as essences (Ahn et al., 2001). From a philosopher's point of view, characterizing the notion of essence is a notoriously difficult task. As Tahko states "philosophers seem to use [essence] in a number of different senses, and even if they do use the notion in the same sense, it is often not quite clear what that sense is" (Tahko, 2018, p. 93), while Mumford contends that neither the commitment to natural kinds nor to intrinsic causal powers are the key commitment of the essentialist theory (Mumford, 2005). In the same vein, characterizing people's biases as 'essentialism' may be an umbrella term for whatever view of the world underlies people's intuitive judgments. I do not think that this a mere terminological issue. If research from cognitive science is to be relevant to constrain metaphysical projects, it is necessary to characterize adequately the structure of the psychological commitments revealed by such research. Is there something better than 'essence' that could capture those psychological commitments?

As we saw in section 2, it is possible to characterize people's commitment to essences as a placeholder (Medin and Ortony, 1989) for whatever it is that underlies the common

properties of a given class. So it is perfectly possible that this underlying "entity" is regarded as something metaphysically different in each case. In particular, it is perfectly possible that it can be sometimes, or for some kinds, understood as an essence in the strongest sense of the word –e.g., an identity-conferring entity that makes objects be the kind of object they are–, but other times as something more flexible –e.g., a causal entity what brings about the object's properties. To characterize this commitment, I will resort to Tahko's notion of unification principle (Tahko, 2022). A unification principle is the narrowest common cause for the clustering of properties in members of natural kinds. Tahko's proposal is that we can appeal to unification principles to vindicate a realist picture of kinds: a kind will be a real/natural kind, insofar as it has an objective unification principle. Armed with this notion, it is possible to construct a new variety of debunking/vindicating argument. In this case, the metaphysical thesis that is the focus of the arguments is not essentialism but natural kind realism. The vindicating version would run as follows:

> *The debunking argument for natural kind realism*
> (i) Realism about natural kind K relies on intuitions that there are objective unification principles for K.
> (ii) Research on psychological essentialism reveals that there are no objective unification principles for K.
> (iii) Realism about natural kind K is thus based on unreliable intuitions.
> hence,
> (iv) Realism about natural kind K is an unreliable view itself.

The vindicating version would be similar, but it would replace (ii) by something like

> (ii*) Research on psychological essentialism reveals that there are objective unification principles for K.

which would lead to the conclusion that natural kind realism for K is a reliable view.

Let me note again that in the debunking/vindicating arguments for essentialism the normative force to assess the reliability of people's intuitions lied outside cognitive science, namely, in the sciences that would make essentialist claims valid or invalid. Cognitive science does not have the power to provide an assessment about the existence of biological or chemical essences. But in the debunking/vindicating arguments for natural kind realism the normative

force comes from cognitive science itself: we are turning to research on psychological essentialism to see if it can provide some unification principles for the natural kinds that people believe to be. How can this be? By looking for these unification principles, so to speak, inside people's minds, i.e., in the mental mechanisms whose nature cognitive science intends to disclose. In the next sections I will outline how this could be done. First, I will elaborate on the notion of unification principle and the problem of mind-independence. Second, I will show how research on psychological essentialism could vindicate or question the existence of unification principles for certain kinds, using the kind 'race' as a putative example.

4. Natural kinds: unification principles and mind-independence

If we want to clarify what sort of commitment to the reality of natural kinds psychological essentialism may allow, it is necessary to examine what counts as a natural kind. This is a widely discussed notion and it is not the purpose of this paper to address it in full. I will limit my focus to a question that plays a central role in the characterization of natural kinds: its alleged mind-independence. Roughly speaking, the idea is that natural kinds are "out there" to be discovered –they are not a construction from our systems of beliefs. Natural kinds are typically contrasted with conventional kinds[5], i.e., kinds that appear as the result of human forms of classification that depend on particular systems of beliefs. The criterion of mind-independence has come under attack because, among other defects, it would leave outside nature certain kinds whose existence depends on human cognitive activities, such as genetically modified organisms, artificially selected organisms, and synthetic chemicals (Ereshefsky, 2018), or certain psychological kinds that may reflect real boundaries in nature despite their mind-dependence. As a replacement of this metaphysical criterion some authors resort to epistemic criteria. For instance, Ereshefsky proposes to replace the criterion of mind-independence by a criterion of defeasibility, so that natural kinds "should be vulnerable to disconfirming evidence" (Ereshefsky, 2018, p. 846). If a classification is mind-dependent but it is useful to investigate the world, then it could count as a natural-kind. As an example of a mind-dependent but epistemically fruitful category he includes race. On the one hand, it is mind-dependent since it depends on people's beliefs about who belongs to a certain race rather than on mind-independent (e.g., biological) properties. On the other hand, it is possible to formulate theories about the effects of being included under a certain racial category. These

[5] Natural vs. conventional does not necessary exhaust all the possibilities. There may be kinds that are not natural but that do not fit the notion of conventional either. Mathematical kinds may be a case in point.

theories can be contrasted with empirical evidence that may defeat the classification, i.e., it might turn out that the socio-economic conditions of people of a certain race are not linked to social discrimination. As an example of a mind-dependent but empirically undefeasible kind, Ereshefsky mentions Khalidi's example of permanent resident as an instance of a conventional kind (Khalidi, 2015). Given that the status of permanent resident could be associated with any number of arbitrary requirements, such as "being able to swim one hundred meters", it is not the sort of category that lends itself to empirical investigation.

However, defeasibility seems to me too weak as a criterion for natural kinds, given that many conventional kinds are susceptible of empirical investigation. Take the notion of permanent resident again. A cursory look at Google Scholar reveals hundreds of studies directed at investigating all sorts of properties of groups of permanent residents. To put but an example, consider an investigation on "which skilled temporary migrants become permanent residents and why" (Khoo et al., 2008), which uses survey data to find patterns of characteristics –such as sex or age– and of reasons – such as liking the country's lifestyle, or getting a better employment– in people who apply for permanent residence. From these data one can formulate hypotheses about other properties that permanent residents may share, and these hypotheses can be certainly defeated by evidence. Conventional kinds have inductive potential as well: just as one may predict certain properties of an individual –Polly– just by knowing that she belongs to a certain natural kind –parrots–, one could predict certain properties of an individual –John Doe– just by knowing that he belongs to a conventional kind –permanent resident. Predictions are defeasible in both cases. To sum up, even if arbitrary conditions, such as swimming capacities, are not defeasible, this does not prevent an empirical investigation of other properties that people obtaining permanent residence may share. Belonging to a conventional kind has empirical consequences that, as in the case of race, can be empirically investigated and defeated. Now, I do not want to conclude that there is no distinction between natural and conventional kinds. I agree that kinds such as permanent resident are not natural kinds, but I do not think that the reasons have to do with their inadequacy to support empirical research. We need to revisit the notion that certain kinds can be found "out there in nature", that is, the notion of mind-independence.

One problem with that notion, as I said above, is that it could be too exclusive. There are putatively natural kinds that can be allowed some forms of mental dependence. The task is to determine which forms are these. Tahko (2022) offers an interesting venue to address the prospects for natural kind realism. His focus is the 'objectivity' of natural kinds, i.e., "the conditions for a natural kind to be real or genuine in a metaphysical sense" (2022, p. 2). To

formulate such conditions, Tahko offers a revisionary account of the mind-independence criterion so that it can encompass all natural or real kinds, including social or mental kinds, and other "higher level" kinds. Tahko's proposal is that kinds are real "insofar as they have an objective unification principle" (2022, p. 2; see also Tahko & Bellazzi, this volume). A unification principle is the narrowest common cause for the clustering of properties in members of natural kinds. A unification principle is not necessarily an essence. There are other possible candidates to do the job, such as causal mechanisms, or laws of nature. These are weaker than essences in the sense that they do not determine membership of a kind. Tahko's proposal prevents us against find an all-encompassing criterion for natural kinds. Instead, the reality of kinds is established in a case-by-case basis. Inasmuch as we can find a certain unification principle for a certain kind, then we have reasons to regard this kind as natural, but we do not need to presuppose that the same type of unification principle will be found in other putative candidates for natural kind.

Tahko contends that searching for unification principles offers a more promising way to account for natural kinds that present some amount of mind-dependence. Take the case of substances that can only exist if synthesized by humans. The narrowest common cause for the properties of such substances is to be found in their chemical composition. The composition itself is independent of the mental states of the humans involved in the activities conducive to its synthesis. To consider a more controversial example, take the issue of psychiatric kinds. Psychiatric conditions, such as depression, are often sustained by the individuals' mental states, such as negative thoughts, or self-schemas. In which conditions could such a kind be regarded as a natural kind? Tahko contends: "It shouldn't matter if, say, the clustering of depression's symptoms is sustained by the depressed individual's psychology. What matters, instead, is whether the general unification principle that underlies this clustering has an objective source, such as a causal mechanism, law of nature, or essence". (2022, p. 19).

Appealing to unification principles could provide a basis to distinguish natural kinds from conventional ones. Take the notion of permanent resident again. The reason why it is a conventional kind does not come from epistemic considerations, but from ontological ones: the reason is that it is unlikely that one can account for whatever properties permanent residents share in terms of a sufficiently narrow common causal mechanism. Moreover, unification principles can also provide the basis to assess when a certain mind-dependent kind does not qualify as a real natural kind. For instance, Tahko argues that the reason why fairy is not a real kind is because properties associated with fairies "are nothing to do with the kind fairy, but rather something to do with the history and psychology of human story telling (…), if there is

an objective UP at play here, it will not be associated with the kind fairy, but rather with some much more general kind concerning human psychology. (…) this type of UP may lack the type of robustness that we have come to associate with natural kinds" (2022, p. 17).

However, there is a problem with this explanatory strategy. When one looks at mind-dependent phenomena, it seems that there are some cases, such as depression, where psychological facts play a constitutive role, and other cases, such as fairy, where psychological facts can be disregarded. To account for this difference, the notion of 'robustness' mentioned in the previous quotation must be spelled out. When can a psychological fact provide a robust basis to constitute (part of) a unification principle? The answer, in my view, lies in the details about that psychological fact. Some psychological phenomena arise from constrained, universal mechanisms that act as the narrow causes for the psychological state; others are based in less constrained and sometimes idiosyncratic mechanisms that are too wide to provide a unification principle. In the following section I will flesh out these ideas by examining the case of race. One may have misgivings about my choice. Would not it be simpler to start with less contentious candidates for natural kinds, such as biological or chemical kinds? My motivation lies precisely in the undeniable mental components in the categorization of races. By examining what cognitive science has to say about such categorization processes I want to assess under what conditions races can be regarded to be "in the head" and, at the same time, have an objective reality. Then these conditions could be extended to other candidates for natural kinds. The upshot will be that there is not a direct general answer and that, as Tahko contends, the prospects to be regarded as a natural kind will have to be determined case-by-case. Some kinds can be mind-dependent in a sense that still allows for unification principles, while other kinds may rely on mental mechanisms that are too unconstrained to provide suitable unification.

5. Psychological essentialism and the unification principles for race

We saw in section 2 that psychological essentialist categorization is not only applied to biological kinds but can be also observed in domains related to human categories that appear more as social than as biological categories. Among the different human categories that have been studied, gender, ethnicity and race consistently appear as those that score higher in essentialism (Prentice and Miller, 2007).[6] As I said, I will focus on human racial categorization mechanisms in an attempt to elucidate the issue of what notion of unification principles we need

[6] Just to have a contrast, the studied category that receives the lowest essentialist score is interests, followed by politics, and appearance.

to search in order to understand how race may, or may not, be regarded as a natural kind. Authors following an epistemic approach to natural kinds tend to offer an affirmative answer about the natural kindness of race. I contended above that epistemic criteria are too permissive to provide an illuminating answer, given that an indefinite number of categories, both natural and conventional, are epistemically significant and have real effects on the world –and race is certainly one of those categories. The question, thus, is not whether racial categories respond to epistemic criteria, but whether they reflect objective distinctions found in nature independently of human minds.

This is a hotly debated question. Many people assume that biological research has already provided a negative answer. Yet there are authors, some of them under the banner of "new biological essentialism" (Devitt, 2023), who argue that natural kinds can be characterized in terms of essences, and that race is one of those kinds. This is not only a philosophical dispute about how to best understand essentialism. Even scientists do not offer a united front in their judgments about what sort of category race is. Working from a sociological perspective, Morning examined the assumption that "everybody knows that race is a social construct" (Morning, 2007). She found that there were differences in the scientists' judgments about the question, with a significant number of them –over one third– holding essentialist beliefs. These differences cannot be explained away saying, for instance, that the essentialist-minded scientists are more misinformed or more liable than their fellow scientists to make mistakes. If we assume that they build their judgments roughly from the same body of evidence, I think that their differences may simply reflect the fact that metaphysical stances are in complex relations with people's beliefs and attitudes. This fact applies both to laypeople and to scholars, including philosophers.

However, apart from social and biological considerations, there are psychological facts to consider. As Cosmides et al. put it: "Race exists in the minds of human beings. But geneticists have failed to discover objective patterns in the world that could easily explain the racial categories that seem so perceptually obvious to adults" (2003, p. 173). If race exists in human minds, what sort of existence is this? Is it a subjective kind, or are there objective, robust unification principles that could unify the kind of race in a suitable way? Tahko contends that there are no such principles for race. Races and witches exhibit the same pattern as fairies, and they must be treated in the same way: "The fact that there is, was, or could be systematic discrimination based on some set of properties that certain groups of people share is not enough, given that this discrimination may be based on a myriad of psychological responses (…) It is precisely because there is no mind-independent unification principle in these cases that we

should discount them, no matter how long lasting or entrenched their causal effects may be" (2022, p. 17). This idea of "a myriad of psychological responses" provides a first specification of the property of "robustness" demanded for unification principles. Psychological responses cannot constitute a unification principle when they are too diverse and originate from an unconstrained variety of sources. Recall that unification demands *narrowest* common causes. An unconstrained class of psychological responses cannot thus provide the appropriate narrowness.

However, when one has a look at the study of the psychological mechanisms that underlie race categorization, one finds reasons to reject the idea that they are based on an unconstrained class of responses. Research about the pattern of recall errors reveals that people encode the race of each individual they encounter, and that this encoding is caused by computational mechanisms whose operation is automatic and mandatory (Cosmides et al., 2003). People report that the primary cue they use to determine race is skin color, and the importance of this cue is independent from the judge's own race or gender (Brown et al., 2007), even if it can be modulated when the judge has additional information (Feliciano, 2016). Skin color has been historically associated to racial distinctions and underlies early attempts to systematize a taxonomy of races (Jablonski, 2021). Yet it could be but one of the features considered in a larger face-recognition system that works on a multidimensional space (Valentine et al., 2016), and there is the hypothesis that race categorization is a byproduct of such a face-recognition system (see Phillips, 2022, for discussion).

These general considerations motivate at the very least the view that the psychological responses underlying people's spontaneous race categorization are, after all, more constrained than their fairy categorizations. They may be based on perceptual mechanisms that are robust and general, more a product of how we are made than of how we are raised. Now, to assess whether those perceptual computational mechanisms are good candidates for unification principles, we should ask if they qualify as the narrowest common cause for the clustering of properties in members of a race. There are reasons to doubt that perceptual mechanisms qualify by themselves: if the mechanisms primarily used for racial categorization are skin color or face-recognition mechanisms, one might well conclude that the kind that is so unified is not the kind of race, but the kind of skin color or the kind of face. How do people jump from these kinds to race? Here is where considerations of psychological essentialism enter the picture. Racial categorization is driven not only by perceptual mechanisms but also by the belief that the individuals that fall under the perceived category share hidden, underlying properties that are responsible for the perceived properties, and that these properties support rich inductive

inferences about further individuals who belong to the class. In other words, if there are unification principles for the kind of race, they are the conjunction of perceptual mechanisms *cum* essentialist beliefs. To see this let us contrast it with cases in which one of the two components is missing.

First, consider, as a case of perceptual mechanisms without the essentialist beliefs, a subgroup of people wearing a tie at a meeting. As this is a perceptually salient feature, they are easy to spot and to tell apart from the rest of people but there is no associated belief about a hidden essence that is responsible for their wearing a tie, or for any other properties of members of the subgroup. Second, consider a case of essentialism with no robust associated perceptual mechanisms. This is the case in many ethnic categories for which people have strong essentialist beliefs yet no straightforward perceptual mechanism to determine category membership. Consider, for instance, how did the Nazis determine who was Jewish: "They used census records, tax returns, synagogue membership lists, parish records (for converted Jews), routine but mandatory police registration forms, the questioning of relatives, and from information provided by neighbors and municipal officials". (https://aboutholocaust.org/en/facts/how-did-the-germans-know-who-was-jewish). The film *The Invisibles* tells us the story of four Jewish who survived Nazi persecution by hiding in plain sight, i.e., by changing their identities so that, with the help of friends who would back their stories, their Jewish background was undetectable by the Nazis. How could they manage to do this? The point is simple: you cannot spot a Jewish person just by looking at her. Contrast this with runaway Black slaves trying to hide their identity.

The point, to repeat, is that the best prospects to find unification principles for the kind of race come from a combination of perceptual mechanisms and essentialist beliefs. There are good reasons to hold that the perceptual mechanisms are robust and constrained enough to count as a narrow common cause for the kind, which is based on how people's physiognomy strikes to us. What about the belief component? Is it also robust and constrained, or can it be the result –as in the case of fairy– of a myriad of psychological responses –i.e., a myriad of culturally-acquired beliefs? The answer is not straightforward because it will depend on how the mechanisms in charge of producing the belief do their job, namely, if they are as robust and constrained as the perceptual mechanisms. To have a glimpse at the space of possibilities, I will summarize the speculations of Cosmides et al. (2003) in this respect. They contend that race encoding cannot be caused by machinery designed by natural selection for that purpose. Instead, they hypothesize that it must be a side-effect of machinery designed for some alternative function. They consider three different alternatives in this respect:

(1) Race encoding could result as a byproduct of domain-general perceptual/correlational systems.

(2) It could be a byproduct of an essentialist inference system that evolved for reasoning about natural kind categories.

(3) It could be a byproduct of computational machinery that evolved for tracking coalitions and alliances.

We could add to this list a fourth alternative that Cosmides et al. apparently overlook:

(4) Racial encoding as a byproduct of *domain-specific* perceptual systems, e.g., face recognition systems.

Each of these alternatives can be backed by theoretical and empirical considerations. For instance, Cosmides et al. favor (3) taking into account experimental evidence about variables that can affect racial judgments, but this is admittedly inconclusive. It is not my goal to review the merits of each alternative in order to assess which one is the most plausible one[7]. Instead, I want to make some general remarks about the consequences that each of them would have for the search of unification principles for the kind of race. First, all of them regard race encoding as a byproduct of another system, not as a dedicated system on its own –none of them posits a domain-specific race categorization system. However, the fact that encoding of a racial kind K is the byproduct of another system S does not rule out the possibility that S counts as a unification principle for K. As the search for a unification principle is the search for the narrowest cause, one has to take into account the processing "distance" between S and K. The closer K is to the function of S (i.e., the fewer intermediate byproducts there are between S and K), the more likely that S can provide a unification principle for K. Conversely, the farther K is from S (i.e., the more intermediate byproducts there are between S and K), the less likely that S can be regarded as the narrowest common cause of K.

Second, domain-specificity plays a pivotal role in assessing how close the encoding is to the processes of the system. Even though the existence of a dedicated system for race

[7] In a recent work that reviews the main theories about racial categorization, Phillips gives some limited credit to the thesis of the face-recognition byproduct while rejecting that there are domain-specific mechanisms for the production of race essentialist beliefs. He contends that these beliefs are largely driven by mechanisms responsible for entitativity perception, and sustains a pluralistic thesis according to which "different ways of thinking about race are driven by distinct mechanisms" (Phillips, 2022, p. 169). Despite this pluralism, I think that the fact that entitativity perception is driven by low-level ensemble-coding mechanisms, may locate his thesis closer to the "narrow common cause" end of the specter. Other authors contend that the tendency to essentialize is an emergent property based on multiple mechanisms and processes, and that psychology itself suffers from an essentialist bias in its search for common underlying mechanisms of the phenomena it studies (Brick et al., 2022).

encoding is dubious, if the system on which this encoding is based is specific for domains that are particularly relevant for racial encoding, then it will be closer to provide the unification principles for race. For instance, if there were a system dedicated to face recognition feeding its outputs straightly to a system dedicated to essentialist beliefs, then the elements constitutive of racial categorization would be narrower.

Applying these ideas to the four alternatives above would yield a rough ordering of them in terms of their prospects to provide a unification principle for racial kinds. At one end, (3) would provide weak unification principles because the causally relevant systems have functions that are very indirectly related to provide racial categorizations. In this account race appears as a result of the workings of mechanisms devoted to track certain social properties, not the properties that act as primary cues for racial categorization, and the beliefs delivered by such mechanisms are not essentialist beliefs but rather more like 'this individual belongs to the in-group or to the out-group'. (1) would provide narrower causes since it hypothesizes a role for perceptual mechanisms. Yet these are still domain-general, not specifically dedicated to racial encoding, so they are still too wide to provide unification principles. (2) would be closer to provide a unification principle since it hypothesizes a system directly devoted to produce essentialist beliefs for natural categories. Yet it does not necessarily posit a system specifically devoted to racial essentialization, given that beliefs about race could be an extension of beliefs about kinds in general. Finally, (4) hypothesizes such a dedicated perceptual system. If this system was coupled with an essentialist-belief producing system such as the one postulated in (2), then we would have the strongest prospects to find unification principles for racial categories and, hence, to conclude that race is a natural kind.

Now, let me be clear about what races would be if they were natural kinds in the sense I am intending. The claim is not that they would be mental kinds. They should not be equated with the mental mechanisms responsible for racial categorization –perceptual mechanisms, belief-producing systems–, or with their mental products –percepts and beliefs. The claim is that they would be real kinds "out there" caused by those mental mechanisms. The kind is not composed by the mechanisms, but by a set of objects: the set of those individuals that our minds classify as belonging to the same racial kind. The objects that fall under the kind of race are special objects inasmuch as the mechanisms bring about the kind by drawing a boundary between those objects and the rest. The same idea can be generalized to ordinary biological and chemical kinds, such as horse or water, to the extent that we can find robust mental mechanisms

that provide unification principles for them[8]. They can be regarded as natural to the extent there is a natural basis for the boundaries that delimit those kinds to appear, only that this basis is to be found in our mental nature. Even though these kinds are mind-dependent, they are objective because the mental mechanisms that bring them about are objective too. They do not depend on our minds, they constitute our minds, and produce some of our entrenched beliefs –such as the beliefs that people categorized in the same racial group share a hidden essence or something close to an essence– sometimes despite our better judgment. These beliefs may well be objectively wrong –i.e., research from other sciences may reveal that there is no hidden essence to reveal– but the existence of the kind does not depend on the correctness of these beliefs –it depends on the universality and the objective existence of these beliefs. Leslie is right in affirming that "our intuitions here reflect only facts about us, not facts about the deep nature of reality" (2013, p. 158). What she misses, perhaps, is that some facts about us belong to the deep nature of reality too, and contribute to reality being the way it is.

Let me turn back to the question of how research on psychological essentialism can contribute to a debunking/vindicating project for some metaphysical thesis. In section 3, I argued that its contribution to debunk/vindicate essentialism is almost irrelevant, given that the strength of the respective arguments lies in what other sciences have to say with respect to essences, e.g., in biology or chemistry. Yet I am arguing now that it can make a significant contribution to debunking/vindicating certain natural kinds, namely, kinds that are revealed to be sustained by the mechanisms studied by cognitive science. In this case, the normative force lies within cognitive science itself: this is the science whose findings and theories we must take into consideration. Our metaphysical conclusions are confined within the limited portion of reality that cognitive science is capable to probe. Moreover, we do not need to depend on debatable reliabilist considerations. In debunking/vindicating a natural kind we do not need to assess whether the proposed mechanisms are reliable or not –as I said, it is likely that the racial essentialist beliefs that they produce are plainly wrong. What we have to assess is if the proposed mechanisms exist and if they are narrow and robust enough to provide a unification principle for the given kind. It is here where the details matter and only cognitive science can provide those details.

I want to end this section with an important remark. Race is a sensitive matter and there is considerable debate regarding how to treat claims about its reality. On the one hand, any

[8] It can also be employed to explain why Jew or permanent resident are not natural kinds, inasmuch as they depend entirely on cultural systems of beliefs, not on facts about our mental machinery. This does not mean they are not real, only that their reality has to be accounted for in a way that differs from the reality of natural kinds.

claims in support of its objective reality can be regarded as fueling some variety of racism. On the other, denying the reality of race has for some people the unwelcome consequence of eroding the basis for positive action to neutralize the pernicious and very real effects of racism in society. The approach I have been developing may suggest a different consequence regarding strategies to deal with racism. If races are classes of individuals that our minds automatically classify as belonging to the same kind and as sharing common "essences", then stating simply "science proves that there are no races" in an attempt to cancel racism is a poor strategy inasmuch as it goes against natural categorization tendencies entrenched in our mental machinery. As Kelly et al. contend, to assess the feasibility of the strategies to cope with racism –e.g., an eliminativist vs. conservationist approach– it is necessary to know the details about how our racial categorization mechanisms work (Kelly et al., 2010; Leslie, 2017). So cognitive science has something to contribute to this normative ethical project as well.

6. Conclusion

Let me revisit the question that opened this chapter: what sort of contribution can cognitive science make to metaphysics through the study of psychological essentialism? My answer is that it cannot do much either to debunk or to vindicate metaphysical essentialism. To this end it would need to show that the psychological mechanisms supporting essentialist beliefs deliver products that are either distorted or largely correct. I argued that they probably have a little of both and that assessing the truth-values of those beliefs is not something that cognitive science can do on their own. Then I turned to a different but related contribution that cognitive science can make: it can reveal the conditions under which certain types of mind-dependence can cause objective, real kinds of objects to appear in nature. They are kinds that depend on robust psychological mechanisms that reveal something deep about how we categorize and reason about categories. Psychological essentialism is neither necessarily a distortion nor a sure road to realism about natural kinds. Its relation to claims about reality is more convoluted. It can explain what is real about the products of our categorization practices by revealing those narrow causes that act as unification principles.

Acknowledgements

This paper is part of Research Project PID2019-108870GB-I00 of the Spanish Ministry for Science and Innovation. I would like to thank the members of the project for their intense

discussion in our seminars, as well as to the participants at the workshop *Kinds of Entities* (University of Granada, Sept. 2022) for their valuable commentaries to the preliminary version I presented. I also wish to thank an anonymous reviewer for this volume.

References

Ahn, W., C. Kalish, S. A. Gelman, D. L. Medin, C. Luhmann, S. Atran, & P. Shafto (2001) Why essences are essential in the psychology of concepts. *Cognition*, 82(1), 59–69.

Atran, S. (1998) Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences* 21: 547–609.

Berent, I., & M. Platt (2021) Essentialist biases toward psychiatric disorders: Brain disorders are presumed innate. *Cognitive Science*, 45(4):e12970.

Bloom, P. (2000) How Children Learn the Meanings of Words. Cambridge (MA): MIT Press.

Braisby, N., B. Franks, and J. Hampton (1996) Essentialism, word use and concepts. *Cognition* 59: 247–274.

Brick, C., B. Hood, V. Ekroll, & L. de-Wit (2022) Illusory Essences: A Bias Holding Back Theorizing in Psychological Science. *Perspectives on Psychological Science* 17(2): 491–506.

Brown, T. D., Dane, F. C., & Durham, M. D. (1998) Perception of race and ethnicity. *Journal of Social Behavior and Personality*, 13 (2): 295–306.

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences*, 7(4), 173-179.

Devitt, M. (2023) *Biological Essentialism*. Oxford University Press.

Ereshefsky, M. (2018) Natural kinds, mind independence, and defeasibility. *Philosophy of Science*, 85(5): 845–856.

Feliciano, C. (2016) Shades of Race: How Phenotype and Observer Characteristics Shape Racial Classification. *American Behavioral Scientist* 60(4): 390–419.

Frugé, C. (2019) Unbunking arguments: A case study in metaphysics and cognitive science. Metaphysics and Cognitive Science. In A. I. Goldman & B. P. McLaughlin, eds. *Metaphysics and Cognitive Science*, Oxford: Oxford University Press, pp. 337–363.

Gelman, S. A. (2003) *The Essential Child*. Oxford: Oxford University Press.

Gelman, S. A. (2013). Artifacts and essentialism. *Review of Philosophy and Psychology*, 4, 449–463.

Gelman, S. A. (2019) What the study of psychological essentialism may reveal about the natural world. In A. I. Goldman & B. P. McLaughlin, eds. *Metaphysics and Cognitive Science*, Oxford: Oxford University Press, pp. 314–333.

Gil-White, F. J. (2001). Are ethnic groups biological "species" to the human brain? Current Anthropology, 42, 515-554.

Goldman, A. I. (2015). Naturalizing metaphysics with the help of cognitive science. In K. Bennett and D. W. Zimmerman, eds., *Oxford Studies in Metaphysics*, vol. 9. New York: Oxford University Press, 171– 213.

Hirschfeld, L. A. (1996) *Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds*. Cambridge (MA): MIT Press.

Jablonski, N. G. (2021) Skin color and race. *American Journal of Physical Anthropology*, 175(2): 437–447.

Kelly, D., E. Machery, & R. Mallon. (2010) Race and racial cognition. In *The Moral Psychology Handbook*, ed. J. Doris and the Moral Psychology Reading Group, 433–472. Oxford: Oxford University Press.

Khalidi, M. A. (2015) Three Kinds of Social Kinds. *Philosophy and Phenomenological Research* 90: 96–112.

Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Leslie, S. J. (2013) Essence and natural kinds: When science meets preschooler intuition. *Oxford Studies in Epistemology*, 4: 108–166.

Leslie, S. (2017). The original sin of cognition: race, prejudice and generalization. *Journal of Philosophy*, 114(8), 393–421.

Lindquist, K. A., Gendron, M., Oosterwijk, S., & Barrett, L. F. (2013). Do people essentialize emotions? Individual differences in emotion essentialism and emotional experience. *Emotion* 13, 629–644.

Livengood, J. & Machery, E. (2007) The Folk Probably Don't Think What You Think They Think: Experiments on Causation by Absence. *Midwest Studies in Philosophy*, XXXI: 107–127.

Machery, E., R. Mallon, S. Nichols, & S. Stich (2004) Semantics, cross-cultural style. *Cognition* 92 (3):1-12

Medin, D. L., & A. Ortony, A. (1989) Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York, NY: Cambridge University Press, pp. 179-195.

Morning, A. (2007) "Everyone knows it's a social construct": Contemporary science and the nature of race. *Sociological Focus* 40(4): 436–454.

Mumford, S. (2005) Kinds, essences, powers. *Ratio* XVIII: 420–436.

Needham, P. (2008). Is water a mixture?—Bridging the distinction between physical and chemical properties. *Studies in History and Philosophy of Science* 39, 66–77.

Neufeld, E. (2022) Psychological essentialism and the structure of concepts. *Philosophy Compass* 17(5), e12823.

Newman, G. E., & Knobe, J. (2019) The essence of essentialism. *Mind & Language* 34: 585–605.

Paul, L. A. (2016) Experience, metaphysics and cognitive science. In J. Sytsma & W. Buckwalter (eds.) *A companion to experimental philosophy* John Wiley, pp. 419–433

Phillips, B. (2022) The roots of racial categorization. *Review of Philosophy and Psychology* 13: 151–175.

Prentice, D. A., & Miller, D. T. (2007) Psychological essentialism of human categories. *Current Directions in Psychological Science* 16(4): 202–206.

Putnam, H. (1975a). "The Meaning of 'Meaning'." In H. Putnam, Mind, language, and reality. Cambridge: Cambridge University Press, pp. 215–71.

Putnam H. (1975b) What is mathematical truth? In *Mathematics, Matter, and Method, volume I of Philosophical Papers*. Cambridge: Cambridge University Press, pp. 60–78.

Rakoczy, H. & Cacchione, T. (2019) Comparative metaphysics: Evolutionary and ontogenetic roots of essentialist thought about objects. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(5), e1497.

Rhodes M., & Moty K. (2020). What is social essentialism and how does it develop? In M. Rhodes (Ed.), *Advances in child development and behavior* Vol. 59. Elsevier, pp. 1–30.

Rhodes, M., Leslie, S., Bianchi, L., & Chalik, L. (2018). The Role of Generic Language in the Early Development of Social Categorization. *Child Development*, 89 (1), 148–155.

Schwitzgebel, E. (2014) The crazyist metaphysics of mind. *Australasian Journal of Philosophy* 92(4): 665–682.

Smith, D. L. (2014) Dehumanization, essentialism, and moral psychology. *Philosophy Compass* 9(11): 814–824.
Strevens, M. (2000) The essentialist aspect of naive theories. *Cognition* 74(2): 149–175.
Tahko, T. E. (2015) Natural kind essentialism revisited. *Mind* 124: 795–822.
Tahko, T. E. (2018) The Epistemology of Essence. In A. Carruth, S.C. Gibb & J. Heil (eds.), *Ontology, Modality, Mind: Themes from the Metaphysics of E. J. Lowe*. OUP, pp. 93-110.
Tahko, T. E. (2022) Natural kinds, mind-independence, and unification principles. *Synthese* 200: 144.
Valentine, T., M.B. Lewis, & P.J. Hills. 2016. Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology* 69 (10): 1996–2019.