Engineering Social Concepts: Labels and the Science of Categorization[*]

Eleonore Neufeld

Department of Philosophy, University of Massachusetts Amherst

June 6, 2023

Abstract

One of the core insights from Eleanor Rosch's work on categorization is that human categorization isn't arbitrary. Instead, two psychological principles constrain possible systems of classification for all human cultures. According to these principles, the task of a category system is to provide maximum information with the least cognitive effort, and the perceived world provides us with structured rather than arbitrary features. In this paper, I show that Rosch's insights give us important resources for making progress on the 'feasibility question' in conceptual engineering: the question of how we can implement conceptual engineering projects in ways that are practically feasible. Specifically, I show that one overlooked upshot of Rosch's work is that naming practices play an extremely important role in the construction of perceived similarities within and dissimilarities between categories, and, correspondingly, the dissemination of social stereotypes that serve as markers between different categories that are otherwise similar. Thus, naming practices will be a crucial constraint for the feasibility of certain ameliorative projects.

1 Introduction

While philosophers have always been, in one way or another, invested in 'conceptual engineering'—the assessment and improvement of our representational devices (Plunkett & Cappelen 2020)—the topic has undergone a significant revival

[*] Draft. This paper is provisionally forthcoming in "Mind, Language, Social Hierarchy" (OUP), edited by Sally Haslanger, Karen Jones, Greg Restall, François Schroeter, and Laura Schroeter. Please cite published version.

in the past 20 years. Haslanger (2000)'s ameliorative analysis of woman has contributed profoundly to philosophers' renewed interest. It inspired a great deal of philosophers to follow suit and give ameliorative analyses of other social concepts (such as, for example, sexual orientation (Dembroff 2016), disability (Barnes 2016), and misogyny (Manne 2017)). However, in more recent years, philosophers started to direct their attention to the fact that many of these proposals have been given in a quite idealized fashion. While the aspiring ameliorators justified their proposals by appealing to certain normative, social, or political goals, they often neglected the question of whether we can practically implement their proposals, given the kind of creatures we are: with a unique psychology, history, and social environment.[1] Conceived this way, the worry goes, conceptual engineering is posed to be a useless enterprise: thinking about conceptual change, no matter how good in theory, won't help us to reach *any* goals if it can't be put into practice. This realization led to the emergence of a new area of inquiry in conceptual engineering, at the center of which lies the *feasibility question*:

- *The Feasibility Question:* how can conceptual engineering be put into practice given contingent factors of our psychology, social environment, and history?

The Feasibility Question centers the kind of non-idealized theorizing that some conceptual engineers have asked for, since it takes as starting point contingent facts about the actual world (Machery 2021). Notably, the Feasibility Question holds particular significance when it comes to *social* categories. Given that our conceptual practices can directly impact the lives of individuals belonging to these social groups, determining whether and, if so, how we can bring about changes in these practices becomes a matter of great urgency with both material and normative implications.[2]

My aim is to inform the Feasibility Question by presenting an important

[1] For some recent literature on this topic, see Pinder (2022); Ritchie (2021); Koslow (2022); Isaac (2020); Machery (2021); Fischer (2020); Neufeld (2023).

[2] For a discussion about the relationship between conceptual engineering and social justice aims, see Catapang Podosky (2022). In line with the Feasibility approach, according to Catapang Podosky, "to promote social justice, conceptual engineering must deliver the following: (i) it needs to be possible to deliberately implement a conceptual engineering proposal in large communities; (ii) it needs to be possible for a conceptual engineering proposal to bring about change to extant social categories; (iii) it needs to be possible to bring a population to adopt a conceptual engineering proposal for the right reasons; and (iv) it needs to be possible to do (i)–(iii) without producing harmful consequences" (Catapang Podosky 2022, p. 159). In this paper, I accept these criteria as background assumptions.

implementation challenge for conceptual engineering projects that aim for the engineered concept to be adopted in ordinary everyday life, by ordinary speakers and thinkers. In essence, the challenge is that certain naming practices play an important role in the construction of perceived similarities and dissimilarities between categories, and, correspondingly, the dissemination of social stereotypes that serve as markers between categories that are otherwise similar. This limitation, I show, follows from the combination of Eleanor Rosch's famous Principles of Categorization (Rosch 1978) and our linguistic naming practices. Therefore, if the goal of conceptual engineers is to improve our representational tools in a way that minimizes our use of stereotypes, the use of labeling practices could make this objective unobtainable. And while several intervention techniques might be proposed to overcome the challenge, none of them, I show, come without problems. Conceptual engineers thus have to think hard about ways to overcome this challenge—either by blocking effects directly, or by offsetting the effects through conceptual changes on other fronts.

Before we start, I want to make a couple of clarifications. In my view, the feasibility focus of this paper pairs best with a 'psychologistic' approach to conceptual engineering.[3] According to this approach,

> "conceptual engineering is concerned with the psychological structures that explain our mental and linguistic behavior [...] to do conceptual engineering is to advocate and implement changes in how people classify things, what inference patterns they are drawn to, and under what circumstances they use particular linguistic expressions" (Koch 2021, p. 1956)[4]

Because the issue of the 'right' approach to conceptual engineering isn't the focus of this chapter, I here assume, rather than argue for, a psychologistic approach to conceptual engineering.[5] In line with this, I also use "concept" in a way that approximates deployment of the term in psychology (cf. Murphy 2004; Machery 2009, 2017; Isaac 2020; Johnston & Leslie 2012). Thus, by "concepts", I here refer

[3] See also Isaac *et al.* (2022) for an insightful overview on the different foci of conceptual engineering projects.

[4] See also Isaac (2020), according to whom conceptual engineering should be concerned with "(re-)modeling the multiply realizable *default* bodies of information that structure our cognitive relationships with reality at large" (Isaac 2020, p. 18, emphasis mine).

[5] For a detailed formulation and thorough defense of a psychologistic practical-aim approach, see Isaac (2020) and Machery (2017). See also Riggs (2019) for related discussion.

to the mental representations of categories that are responsible for our ability to categorize and perform inductions in a host of reasoning processes. Philosophers often prefer to use the notion of "conceptions" for these kinds of cognitive structures (Burge 1993).[6]

Having settled terminological questions, I can now present the plan for this chapter. In §2, I present Rosch (1978)'s Principles of Categorization and other theoretical background on the cognitive science of concepts. In §3, I show that the Principles imply that certain labeling practices can result in the construction of stereotypes in ways that are often overlooked. This generates an important implementation challenge for ameliorative projects. In §4, I consider possible solutions to the challenge, and show that none of them come without problems. I close, in §5, by noting avenues for future research.

2 Rosch on Concepts and Categorization

2.1 *The Principles*[7]

Jorge Luis Borges' fictional taxonomy in *Celestial Emporium of Benevolent Knowledge* famously classifies animals into the embalmed ones, those that are trained, those that belong to the emperor, suckling pigs, mermaids, fabulous ones, stray dogs, and many others (Borges 1937). What's remarkable about this taxonomy is that it does not exist—we have no lexicalized terms that pick those categories out. Why not? In one of her many seminal papers on the psychology of categorization, Eleanor Rosch provides us with an explanation (Rosch 1978). Human categorization is not arbitrary. Instead, two psychological principles constrain possible systems of classification for all human cultures. Here is the first principle:

- *Principle of Cognitive Economy* The task of a category system is to provide maximum information with the least cognitive effort.

[6] For discussions of the relation between 'concepts' in the psychological and 'concepts' in the philosophical sense, see Löhr (2020); Johnston & Leslie (2019); Machery (2009); Nefdt (2021). Note that even if you read this chapter as being about the change of people's *conceptions*, it still falls squarely under conceptual engineering. As mentioned at the outset, standard characterizations of conceptual engineering treat it as the assessment and improvement of our representational devices (Plunkett & Cappelen 2020; Isaac *et al.* 2022). 'Conceptions' of certain categories are clearly representational devices and directly responsible for our conceptual practices, such as our classificatory and inferential behavior.

[7] This subsection draws heavily from Neufeld (2023).

This principle strikes a compromise between two distinct pressures driving our cognitive system: to get as much information as possible from an act of categorization, and to preserve finite cognitive resources. A natural extension of this principle is that the concepts most useful and basic for us are those that have a high degree of *similarity* and *distinctiveness*. 'Similarity' describes the probability that a certain feature is present, given that something is an instance of a category: $p(Feature|Category)$. 'Distinctiveness' (sometimes referred to as 'cue-validity') describes the probability that an instance belongs to a category, given that it has a certain feature: $p(Category|Feature)$. Concepts with these attributes will allow us to maximize both informativeness and ease of categorization.[8]

It is useful to illustrate this with an example. Consider the category *dog*. When we categorize something under the concept DOG, we can draw many useful inferences about it: that it has fur, four legs, a heart, lives with humans, etc. The reason we can draw this many inferences is because members of the category are quite similar to each other. In other words, the category has high within-category similarity. At the same time, we also preserve cognitive resources because members and *non-members* of the category are very dissimilar to each other—i.e., DOG is associated with many distinctive features. Consider now the contrast between DOG and GIRAFFE. Because each category is associated with very distinct features, you don't have to run through a long feature search to tell them apart. Once you detect that something barks, you can reasonably infer it's a dog; once you detect that something has a very long neck, you can reasonably infer it's a giraffe.

Let's turn to Rosch's second principle:

- *Principle of Perceived World Structure* The perceived world comes as structured information rather than as arbitrary or unpredictable features.

Behind the principle is the simple truism that some properties co-occur with other properties more often than with others, and the perceived world reflects those bundles of co-occurring features. Here's a simple example: Manes usually co-occur with lion bodies, and they rarely co-occur with taxis. Thus, information we get from the perceived world is rich and not unpredictable.

To see how these principles help explain the fact that Borges' fictional taxonomy is only fictional, suppose entities x and y both resemble flies from a distance, so we classify them under the concept RESEMBLING FLIES FROM A DISTANCE. Assuming the

[8] For computational models, see, e.g., Jones (1983); Gosselin & Schyns (2001); Tversky (1977).

classification is correct, what else can we predict about them? Not very much. Given the sort of world we live in, x and y are likely to share few additional properties (other than looking like flies from a distance) with each other. Furthermore, their features are not distinctive. The flies from a distance could also be stones, or bats, or birds, or airplanes, and so on. In contrast, suppose x and y both look like lions, so you classify them under LION. Assuming the classification is correct, what else can we predict about them? Given the structure of the world, we can predict quite a bit: that they probably have a heart, fur, are mammals, hunt, live in Africa, are carnivores, have tails, and more.

2.2 *The Levels*

Importantly, as Rosch points out in the same paper (and on many other occasions), the Principles of Categorization (henceforth: "Principles") also imply that not all ways of categorizing the world will be equally useful. To see this, it is useful to introduce an important notion of the cognitive science of concepts: the *hierarchy of categorization*.

The first cognitive scientist who drew systematic attention to the fact that many common concepts are embedded in a hierarchical organization was Roger Brown. In his 1958 paper, he pointed out that

> When such an object is named for a very young child how is it called? It may be named *money* or *dime* but probably not *metal object*, *thing*, *1952 dime*, or *particular 1952 dime*. The dog out on the lawn is not only a *dog* but is also a *boxer*, a *quadruped*, an *animate being*; it is the *landlord's dog*, named *Prince*. How will it be identified for a child? Sometimes it will be called a *dog*, sometimes *Prince*, less often a *boxer*, and almost never a *quadruped*, or *animate being*. Listening to many adults name things for many children, I find that their choices are quite uniform and that I can anticipate them from my own inclinations. (Brown 1958, p. 14)

This early insight of Brown's became an important objects of systematic study within the cognitive science of concepts (cf. Rosch *et al.* 1976; Rosch 1978; Murphy & Lassaline 1997; Hampton 1982; Markman & Wisniewski 1997; Murphy 2004). Notably, Eleanor Rosch developed, refined, and studied the hierarchy of categorization systematically. The general insight, also reflected it Brown's quote above, is that

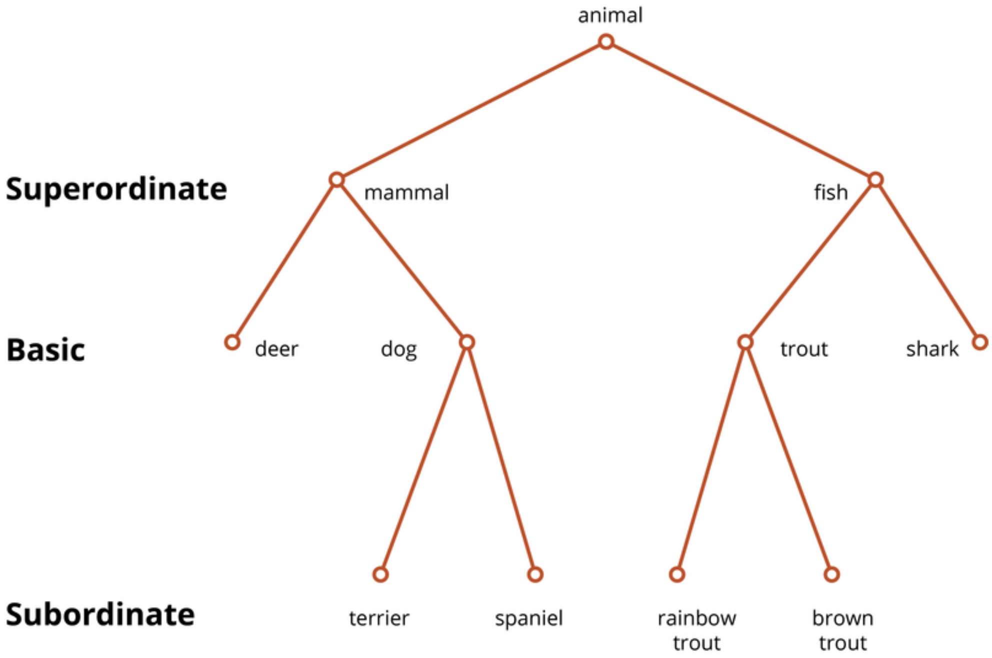

Figure 1: Simplified illustration of hierarchically organized categories from Murphy (2023), licensed under CC BY-NC-SA 4.0.

many concepts are organized in a hierarchical taxonomy according to their degree of inclusiveness. As you can see in fig. 1, you can categorize a thing according to different levels of inclusivity, where the top levels are most inclusive, and the bottom levels are least inclusive but most specific. For example, it is possible to categorize something as cocker spaniel, dog, mammal, or animal, where cocker spaniel is the most specific category, and animal the most inclusive.[9] Within the cognitive science of concepts, it is convention to call levels of high inclusiveness (such as *mammal* or *animal*) "superordinate levels", levels of high specificity and low inclusiveness (e.g. *terrier* or *spaniel*) "subordinate levels", and the middle level of specificity and inclusiveness the "basic level" of categorization. Correspondingly, lexical concepts in these levels are called basic level concepts, superordinate level concepts, and so on.

Notably, the categories in the hierarchy seem to play a special role in communication and cognition. When asked to categorize an object in a neutral setting, you will categorize it as dog, animal, or poodle, rather than as a tailed thing or

[9] Of course, it is possible to categorize in a way that's even more general or specific.

something that walks the streets of Holyoke (cf. Murphy & Lassaline 1997, p. 94). But even among these levels, one level of categorization is particularly privileged in cognition and communication:

> Of all the possible categories in a hierarchy to which a concept belongs, a middle level of specificity, the basic level, is the most natural, preferred level at which to conceptually carve up the world. The basic level can be seen as a compromise between the accuracy of classification at a maximally general level and the predictive power of a maximally specific level. (Murphy & Lassaline 1997, p. 100)

The quote by Murphy & Lassaline nicely crystallizes two key points about basic level concepts.

First, the basic level is the *preferred*, *most natural* level to conceptually carve up the world. For most purposes, it is most natural to categorize something as car, rather than VW Passat or vehicle; as sofa rather than as chesterfield sofa or furniture; as lion rather than mammal or Asiatic lion; and so on.[10] Echoing the earlier quote from Brown (1958), when asking your table neighbor for salt, odds are you will ask them to pass the *salt*, not the sea salt or the mineral. All this isn't only supported by intuition, but has been underscored by a host of evidence involving all sorts of cognitive tasks. For example, in naming tasks of pictures, people predominantly use the basic level term (Jolicoeur *et al.* 1984; Tanaka & Taylor 1991; Rosch *et al.* 1976). In free naming paradigms in which subjects were asked to label pictures with the first name that came to mind, 1595 names used were at the basic level, while 14 subordinates and only one superordinate term were used (Rosch *et al.* 1976). People are also faster at categorizing objects at the basic level in verification tasks (Rosch *et al.* 1976; Murphy & Smith 1982; Murphy & Brownell 1985; Mervis & Rosch 1981), and within development, basic level terms come first in language acquisition (Smith 1926; Rosch *et al.* 1976; Callanan 1985; Mervis & Crisafi 1982; Horton & Markman 1980).

Most important for the focus of this paper, basic level *terms* also play a special discursive role. In fact, an often-used *heuristics* to find out about "basic-level-ness"

[10] Similarly, Brown notes: "[...] both children and adults make some distinction among these various names. The name of a thing, the one that tells what it *"really"* is, is the name that constitutes the referent as it needs to be constituted for most purposes. The other names represent possible recategorizations useful for one or another purpose. We are even likely to feel that these recategorizations are acts of imagination, whereas the major categorization is a kind of passive recognition of the *true character* of the referent" (Brown 1958, p. 17, my emphasis).

is to look at the role of terms in discourse: "[t]he simplest way to identify an object's basic-level category is to discover how it would be *labeled* in a neutral situation" (Murphy 2023, emphasis mine). Correspondingly, analyzing corpora of oral descriptions of movies, Downing (1977) found that people used terms at the basic level most frequently. Similarly, analyzing printed text, Wisniewski & Murphy (1989) found that basic level terms are used most frequently when describing single objects. There are several other evidential markers that signal basic-level status in language, such as word length, monolexicality, frequency, and, in some languages, grammatical gender and the presence of semantic radicals (Zubin & Köpcke 1986; Murphy & Lassaline 1997; Rosch *et al.* 1976; Callanan 1985; Wang *et al.* 2018).

Second, the *reason* why the basic level of categorization is cognitively privileged is closely connected to the Principles. Notice that each level of categorization comes with important trade-offs. The superordinate level has high distinctiveness or cue validity. The category *furniture* is very different from the category *animal*; they barely have any features in common. But at the same time, superordinate categories don't have much within-group similarity. Members of the category *animal* are very different from each other: jelly fish, koalas, and butterflies don't have that much in common. Due to this lack of similarity, we can't draw *that* many inferences from knowing that something is an animal or a piece of furniture. This goes against what is prescribed by the Principles.

We *maximize* informativeness at the most specific level, the subordinate level. This is because categories on this level are most specific and therefore have most within-group similarity: Bengal tigers are even more similar to each other than tigers. However, we gain this informativeness at the expense of efficiency. In fact, we only gain very few inferences by this increase in specificity (e.g., we only know a little bit more about a thing if we know it's a Bengal tiger vs. a tiger),[11] but need much more cognitive effort to tell members of the category apart from other subordinate categories, because they're quite similar to each other (e.g., a Bengal tiger vs. an Indochinese tiger). At the same time, categorizing things at the subordinate level also increases our chances of getting things *wrong*, simply because different categories on this level share so many features and are easy to mistake for one another.

[11] Correspondingly, in feature-listing tasks, participants list substantially more features for basic level categories than superordinate categories, and only slightly more for subordinate categories (Rosch 1978).

The basic level of categorization, however, seems to optimize *both* informativeness and cognitive efficiency, because it stands at the optimum of within-group similarity and between-group distinctiveness (Markman & Wisniewski 1997; Murphy & Brownell 1985; Murphy & Lassaline 1997; Murphy 2004; Rosch *et al.* 1976; Rosch 1978). Consider our examples from before: it is easy to distinguish dogs from giraffes because of their distinctive features, while we can at the same time infer a lot from knowing that something is a dog. This is simply because the basic level is the most inclusive level of categorization at which objects in the class share a high number of common and distinctive attributes, so high similarity and distinctiveness are 'baked' into the level.

In sum, then, the Principles do considerable explanatory work for a theory of categorization: they help us understand why we form certain category taxonomies rather than others, and why even among those, a certain level of categorization is privileged: the basic level. In the next section, we'll extract lessons from the generalizations we've reviewed in this section for the prospects of ameliorating social concepts.

3 Labels, Meta-Evidence, and the Principles

One of the main upshots from the last section is that the way we categorize the world—including the social world—*makes sense* given the two principles of categorization. Because efficiency constraints underlie our category system, and because our perceived world presents certain features and not others as co-occurring, our category system contains certain concepts rather than others. And even among those, the ones at the 'basic level'—i.e., the ones that hit the 'sweet spot' between similarity and distinctiveness—are particularly privileged. Rosch's ingenious work contains invaluable insights for the psychological feasibility of conceptual engineering projects. In the remainder of this paper, I will present and discuss in detail one of these insights: naming practices will be a crucial constraint for the feasibility of certain ameliorative projects.

In the same paper that spells out the theoretical basis of the Principles, Rosch (1978) also notes that:

> One influence on how attributes will be defined by humans is clearly *the category system already existent in the culture at a given time*. Thus,

> our segmentation of a bird's body such that there is an attribute called "wings" may be influenced not only by perceptual factors such as the gestalt laws of form that would lead us to consider the wings as a separate part [...] but also by the fact that at present we already have a cultural and linguistic category called "birds." (Rosch 1978, p. 313, my emphasis)

Rosch presents this point as a rather parenthetical remark. Correspondingly, she leaves open *how* or *why* the category system already in place in a culture has the "influence" she's alluding to. The answer, I think, lies in the Principles. Although not directly apparent, the Principles assign a key role to public systems of classifications for our common psychology of categorization. The Principles, recall, claim that our cognitive category system has at least two properties: it is economical in that it optimizes the gains and costs of an act of categorization via diagnosticity and similarity, and in doing so, it makes use of statistical structures contained in the perceived world. If that's the case, it means that a cognitive agent can *extract important information from the lexicalized category system that's in use in their community*. Namely, that it is economical in the above sense—i.e., that the members of a category have properties that are similar and/or distinctive—and that it mirrors statistical regularities from the perceived world. In other words, *given* that we adhere to Rosch's principles of categorization, the conceptual practices we adopt as a community contain important meta-evidence about category members and (alleged) non-members that go over and above information we directly associate with the category in question.

As a result of this, a given lexicalized conceptual practice will then indirectly communicate that the category's features score high on distinctiveness and similarity metrics, and that the features mirror the statistical structure of the perceived world. So by classifying two people with a label L, we don't only communicate that they are both L, but also that they are sufficiently similar in important respects as to warrant classification under the same concept. This has important attention-guiding consequences, insofar as it primes agents to search for similarities underlying the co-classification, even if those aren't immediately obvious.[12] Similarly, by classifying two people through distinct labels, we communicate that they are sufficiently dissimilar in important respects as to warrant classification

[12] See Whiteley (2023) for a discussion of attention to social identity.

under different concepts. This primes agents to search for dissimilarities in order to make sense of the classification—and if there aren't many dissimilarities, to exaggerate the statistical weight of the ones there are. As a result, naming practices play an extremely important role in the construction of perception of similarities and dissimilarities, and, correspondingly, the dissemination of social stereotypes that serve as markers between different categories that are otherwise similar.

This issue will become particularly acute when it comes to *basic level* concepts and their corresponding labels. Recall that basic level concepts hit the "sweet spot" between distinctiveness and similarity. Thus, usage of category labels with the 'basic level' status gives rise to exactly those meta-inferences about high intra-group similarity and high inter-group difference. Someone might object that an agent can't independently know which categories have the 'basic level' status in a conceptual community. However, recall that there are numerous *cues* that serve as evidential signals for basic level status in a conceptual community. Among other things, labels for basic level categories appear more frequently in written and verbal discourse, have shorter word length, and are more frequently monolexemic. In addition, a cognitive agent might simply register (potentially subconsciously) which level of specification is used in labeling practices as *default*. Applied to *social* categories, when a social label has these properties, an agent will have evidence that they are dealing with a basic level category—i.e., a category with sufficient ingroup similarity and outgroup distinctiveness to warrant preferring this term over others.[13]

To make matters worse, once a basic-level term has had the communicative-inductive impact outlined in this section, it might interact with pre-existent cognitive biases and lead to *further* exaggeration of the feature weight we associate with distinctive features. In their ingenious work of stereotypes, Bordalo *et al.* (2016) model them as a consequence of Tversky and Kahneman's representativeness heuristics (Tversky & Kahneman 1983). An "attribute is representative of a class if it is very diagnostic; that is, the relative frequency of this attribute is much higher in that class than in the relevant reference class" (Tversky & Kahneman 1983). Formally, a feature f is representative for group G relative to a comparison group

[13] These insights pair well with empirical data on social categorization. There's a wealth of evidence that social basic-level categorization reduces perceived variability within social groups (Linville *et al.* 1989; Messick & Mackie 1989; Mullen & Hu 1989; Fiske & Taylor 2013; Devos *et al.* 1996). And, as predicted by our account, naming practices can have stark effects on both perceived group homogeneity and distinctiveness (Gelman & Heyman 1999; Carnaghi *et al.* 2008).

$\neg G$ if it scores high on the likelihood ratio in (1) (Gennaioli & Shleifer 2010):

(1) $$\frac{Pr(f|G)}{Pr(f|\neg G)}$$

So far in this section, I have argued that basic-level terms can prime us to believe that a group G has features f that are representative in the sense of (1), even if it doesn't. But as Bordalo *et al.* (2016) demonstrate, once features are represented as representative, they "come to mind first and so are overweighted in judgements. Predictions about G are then made under a distorted distribution, or stereotype, that overweights representative types" (p. 1755). Bordalo *et al.* (2016) offer an ingenious (and empirically tested) formal model of the distorted *recall* that results from representativeness. As a result of the distorted recall, differences between groups get amplified further, resulting in a highly distorted distribution. Assuming this model is correct, in our cases, we have *two*, not one, stages of overweighting. In the first stage, we increase the distinctiveness of categories in order to make sense of the fact that it is, presumably, as basic level category. In the second stage, we increase its weight further due to distorted recall.

Let us now work through a couple of examples that illustrate how, concretely, the pre-existence of basic level labels for social categories might affect our conceptualization of them. Consider national basic level categories. There's a clear sense in which members across categories such as *Americans, Turks, Germans, Peruvians*, and so on aren't very different from one another. That isn't to say there aren't any differences—e.g., there will at least be a common and distinctive difference in nationalities—but when we zoom in on those categories, the differences seem to be of entirely different magnitude when contrasted with basic level categories from, say, the animal kingdom. The number of differentiating features for the categories *elephant* and *squirrel* far exceed the ones for *Peruvian* and *American*. At the same time, in our linguistic communities, we pick out those categories via the basic level terms "American" and "Peruvian". This linguistic practice (indirectly) communicates that these categories are associated with *distinct features that are common* among the category. In order to make sense of the classificatory practice, a cognitive agent will, as a result, be primed to look for distinguishing features, and infer that the distinguishing features that do exist carry higher statistical weight than was initially assumed.

This explanation is directly relevant to open questions about the existence and formation of stereotypes. In a well-known series of studies sometimes referred to as the 'Princeton Triology' (Katz & Braly 1933; Gilbert 1951; Karlins *et al.* 1969), participants were asked to list as many attributes as possible that best characterize different national and/or ethnic groups: Germans, Italians, African Americans, Irish, English, Jews, Americans, Chinese, Japanese, and Turks. In the latest follow-up of the series (Madon *et al.* 2001), students reported that different nationalities/ethnic groups were associated with substantially different stereotypes. For example, while Germans were characterized as intelligent and industrious, Turks were described as 'extremely nationalistic'. It seems hard to explain, however, *why* we harbor these stereotypes. They don't track the statistical structure of the perceived world, and don't maximize information we get out of an act of categorization. For example, the probability of someone being German, given that they're intelligent, is fairly low, and intelligence performance is normally distributed with the usual median, meaning the vast majority of Germans won't be extraordinarily intelligent. But given that we appeal to these categories via basic-level labels, we can, at least partially, explain why we associate the categories with stereotypical features that don't maximize utility in the way predicted by the Principles. By having in our public lexicon basic-level labels for nationalities such as "German", "Turk", "American", etc., we communicate that these groups have high ingroup-similarity and intergroup-difference. As a member of this linguistic community with only limited direct access to information about the features of all social groups, the linguistic practice will serve as a valuable piece of meta-evidence. Even if you don't know *which* features are distinctively prevalent in a given group, the label serves as a cue *that* the property structure of the group is such that it has features that are highly distinctive and common. So when encountering a candidate diagnostic feature—such as industriousness or nationalism—we will represent it as more prominent within the group, and less prevalent in other groups, as to make sense of the basic level classificatory practice.

Another example that illustrates how the interface between the Principles and extant linguistic practices might affect our representation of social categories is the debate surrounding the term "woman". Some gender-critical philosophers have argued that the term "woman" is correctly applied only to adult human females (Byrne 2020; Stock 2018). Note that in general, "woman" is a basic-level social label that communicates, again, that members of the group have high ingroup-similarity and intergroup-difference.[14] This restrictive usage of the term might have indirect consequences on the formation of stereotypes associated with trans women. Recall that by classifying people through *distinct* labels, we communicate that they are sufficiently dissimilar in important respects as to warrant classification under different concepts. If the basic-level term "woman" won't be applied to trans women, members of a linguistic community might try to make sense of the practice by looking for common *differences* between those that the term "woman" is applied to and trans women. If they don't find any obvious differences that warrant lack of inclusion in the same basic-level category, cognitive agents will possibly exaggerate the statistical weight of the differences they do spot. In this way, linguistic activism for restrictive application of the label "woman" by gender-critical philosophers might effectively establish enabling conditions for the formation of problematic stereotypes of trans women.[15]

Summing up, we have seen that naming practices can play an extremely important role in the construction of perception of similarities and dissimilarities, and, correspondingly, the dissemination of social stereotypes that serve as markers between different categories that are otherwise very similar. Importantly, while I argued that that naming practices can give rise to stereotypes in the way I outlined, it is important to emphasize that they are in *no* way the only, or even the major, source of social stereotyping. As I emphasize in other work (Neufeld 2019, 2020b, 2023, 2022), there are multiple sources of stereotypes, and multiple cognitive structures subsumable under the category *stereotypes*.[16] I here just highlight one—maybe underappreciated—mechanism of stereotype formation that becomes clear if we apply Rosch's Principles to our naming practices. This is because, as I will show in the next section, this insight poses important challenges for the prospects of conceptual engineering projects.

[14] See Deaux *et al.* (1985) and Harper & Schoeman (2003) for evidence that woman is a basic level concept. Interestingly, Deaux *et al.* (1985) tested the hypothesis that gender categories such as woman and man are *not* a basic-level categories, on the grounds that they seem so broad that it's unlikely for them to occupy a middle-level of inclusivity. Contrary to their hypothesis, and consistent with the view that gender concepts *are* basic, they found that these gender categories are associated with rich category associations.

[15] Note that my claim is not that inclusive application of the basic-level term "woman" would eradicate other stereotypes associated with women. My claim is only that usage consistent with gender-critical views could potentially give rise to further stereotypes of women; specifically, stereotypes of trans women.

[16] For other recent work on stereotypes, see Madva & Brownstein (2018); Puddifoot (2021); Johnson (2020); Bosse (2022); Westra (2019); Del Pinal *et al.* (2017); Beeghly (2015).

4 Do We Have Solutions?

In the last section, we have seen that naming practices play an important role in the construction of perception of similarities and dissimilarities, and, correspondingly, the dissemination of social stereotypes that serve as markers between different categories that are otherwise very similar. This insight poses an important challenge for ameliorative projects of social concepts: specifically, labeling practices might limit the conceptual improvement that's feasible to do. Suppose an ameliorative proposal aims at improving representational devices that give rise to stereotypes about a given social group. Insofar as we talk about the group via a basic level label, the project is at risk to fail. Are there principled ways to overcome the challenge? In this section, I consider three potential strategies: expertise, complex linguistic constructions, and top-down adjustment. We will see that each approach comes with substantial problems. Though the proposed strategies may not cover all possible solutions, this discussion underscores the difficulty of addressing the implementation challenge. Thus, it warrants considerable attention in the realm of conceptual engineering research.

4.1 Expertise

A first possible strategy would be to take advantage of the *flexibility* of the hierarchy of categorization. Adequately outlining this strategy require us to branch out, again, into some psychology. In previous sections, I sometimes made it sound as if what we represent as basic, subordinate, and superordinate (and so on) is fixed. But this isn't true. Although we can certainly make claims about the way a certain category will generally be represented (e.g., that car and bike are generally basic-level concepts, and vehicle a superordinate concept), what we represent and treat as basic-level category can be quite flexible. Specifically, what we called 'basic level' in this chapter often changes as a function of expertise we have with a class of objects (or animals, people, etc.) (Murphy 2004; Medin *et al.* 2000; Tanaka & Taylor 1991; Johnson & Mervis 1997; Clarke & Tyler 2015; Berlin 2014). What this means is that the concept we directly and immediately apply to an object is variant, depending on factors such as familiarity or relevance for the tasks we are steadily confronted with. Imagine you are a Volkswagen car seller. Do you think car would be your preferred mode of categorization? Probably not. After some time in your position, you would certainly start recognizing a Tiguan directly *as* a Tiguan, a Golf *as* a Golf, a Passat directly *as* a Passat, and so on. That is, what you previously might have treated as subordinate level concept has gained the status of 'basic level concept' due to your expertise. You've become so familiar with the relevant categories that you are now able to pay attention to many details that make the relevant, more fine-grained categories *distinctive*.

In the last section, I pointed out that our *default* 'level' of carving up social groups seems to differ from the default level in which we carve up other categories, such as animals or artifacts. Many salient social *basic* level categories—e.g., different ethnic groups—seem to have the degree of inter-group similarity that groups have we usually treat as *subordinate*-level categories, such as *Bengal tiger* and *Siberian tiger*. Thus, it seems that we encode many social concepts as basic although their similarity profile would make it more apt to treat them as subordinate level concepts. All of this would be unproblematic if we had substantial expertise with the corresponding social categories. And in fact, it often seems to be the case that we *are* experts when it comes to the social world. Social features have higher salience and greater relevance in many contexts (Neufeld *et al.* 2016; Neufeld 2020a). Thus, it shouldn't be surprising that our basic level is generally more fine-grained for social categories. However, in many cases, we're *not* experts when it comes to a social category in question. These cases should be particularly prone to elicit the problematic effects we discussed in the last section. Imagine being forced to, *by default*, categorize ants in a very fine-grained way, while simultaneously not knowing very much about ants: as Carpenter ant, Argentine ant, Sugar ant, and so on. Predictably, you would make lots of mistakes, both in categorizing, and in inferring traits based on category membership.[17] This is precisely the situation we might find ourselves in when it comes to many social categories. But the realization that our default level of social categorization doesn't align with the required expertise might also open an avenue for feasible conceptual amelioration: we can *align* our level of expertise with our used level of categorization.

In fact, the realization that we don't have the required social 'expertise' to use many social categories at the level we do dovetails nicely with the so-called 'contact hypothesis', according to which intergroup contact is conducive to the reduction of prejudice (Allport 1954; Dovidio *et al.* 2003; Pettigrew & Tropp 2006; McKeown & Dixon 2017). Presumably, intergroup contact will precisely provide the level of

[17] Correspondingly, we are generally prone to more errors when categorizing at the subordinate level without expertise (Murphy & Lassaline 1997).

'expertise' required to supply attentional training and increase accuracy of feature distribution and differentiation. So here, we might have an instance of feasible amelioration supported by the cognitive science of intergroup relations: If we align actual expertise with the level of specificity at which we talk about social categories, naming practices might cease to have the undesirable effects outlined in the last section. And we can achieve this by simply increasing people's intergroup contact.

However, while it sounds initially promising, there are several problems with this approach. While empirical results from the last decades seem to lend general support to the basic gist of the hypothesis (Zingora *et al.* 2021; Kenworthy *et al.* 2005; Pettigrew *et al.* 2011), several critics have pointed at methodological limitations of the research (Bertrand & Duflo 2017; Dixon *et al.* 2005). In fact, as Dixon *et al.* (2005) point out, several conditions have to operate in concert in order for intergroup contact to have positive effects, and it is highly unlikely for these conditions to take place simultaneously. If they don't, intergroup contact can backfire, and can lead to increased stereotyping, as is often the case in contexts of immigration and refugee crises (Hopkins 2010). Thus, aiming for aligning basic-level naming practices with expertise seems to create a feasibility problem all of its own.

4.2 *"person from..."*

In the last section, our strategy was to find a way to align the degree of specificity indirectly communicated via the labels we use to pick out social groups with the group's actual level of specificity. Another way of following this approach, we might think, involves promoting the use of linguistic vehicles that don't *signal* basic level status. As mentioned earlier, certain formal properties of labels can reliably signal basic level status in our linguistic community. Thus, one way towards alignment might then consist in choosing constructions that *don't* signal basic level status, such as longer, more complex phrases. For example, "Germans", "Turks", "man", "Mexicans", "Christians", "Jews", "blacks", etc. could be re-described as "*person* from Germany", "person who identifies as man", "person with Christian beliefs", etc. The hope would be that these linguistic constructions make clear that we are dealing with a subclass of a more general group (i.e., *person*), making it more likely to assign the group subordinate, rather than basic level status. As a result, we wouldn't provoke members of our linguistic community to look for many highly distinctive and common features to rationalize the linguistic practice of picking out social groups via basic level labels.

An advocate might try to bolster this approach by appealing to work in pragmatics. For example, Levinson (2000) argued that using non-lexicalized expressions can give rise to M-implicatures. M-implicatures are implicatures that are generated via the M-Principles: "Indicate abnormal, nonstereotypical situations by using marked expressions that contrast with those you would use to describe the corresponding normal, stereotypical situations" (Levinson 2000, p. 136). To illustrate the principle, consider (2) and (3):

(2) Kirsten killed Colleen.

(3) Kirsten caused Colleen to die.

According to Levinson, due to the availability of a different, lexicalized morpheme in the linguistic community, (3) becomes marked and pragmatically signals that Kirsten did not kill Colleen, but caused her to die in some non-conventional way. By extension, constructions like "person from Germany" might similarly trigger an M-implicature: a hearer might infer that they are *not* dealing with a distinctive, basic category with high inductive potential.

In fact, a similar proposal has famously been put forward by Sarah-Jane Leslie (2017). Noting the essentializing effects of basic level nouns (see also Neufeld 2019, 2022), she suggested to avoid their deployment in favor of more complex adverbial constructions, especially in the context of generic language. However, Leslie (2017)'s proposal didn't come without criticism. Among others, Jennifer Saul (2017) cautioned against the linguistic reform proposed by Leslie:

> Leslie suggests that we should try to eschew 'labels' and opt instead for 'descriptions'. And it might be true that when we initially replace 'Muslim' with 'person who follows Islam' (a suggestion that she makes), we'll be slower to ascribe an essence. But soon that phrase will simply be a label, and function as one. [...] And this is not just a speculative point. It is worth reflecting also on how notably unsuccessful it was to replace the noun 'moron' with the descriptive phrase 'mentally retarded person'. The more recent terms 'special needs' and 'person with special needs' also provide a revealing case study. Indeed, the drive to label groups with noun phrases has led the noun phrase 'special needs' to be used as an adjective in 'special needs children'. Even when it's (initially) ungrammatical, we will find a way to form the easy noun phrases that

> facilitate essentialising. Reflecting on cases like these should give one pause about the efficacy of attempting to reduce essentialising through this sort of linguistic reform. (Saul 2017, p. 14)

It seems like a similar concern would extend to the approach under consideration. If we use "person from Germany" instead of "Germans" every time we would normally use the latter, the use might become so frequent that we treat it as basic level after all. As a result, then, "people from Germany / Peru / Turkey / X" will just stand in for a group that's supposedly distinctive and highly predictive, and lead to the same problem of stereotype generation. So although the idea to simply use more complex linguistic constructions in order to signal subordinate status and block inferences about highly distributed differences might sound attractive on its face, Saul's concerns show that it might not be a promising strategy after all.

Before continuing, it's worth pointing to another problem with the present approach. According to the strategy under consideration, we should avoid the deployment of (basic level) labels, and instead use complex constructions that signal subordinate status. But depending on background conditions, basic level labels can also have important positive effects. Especially in conditions in which labeling of minorities by dominant groups is rampant, *self-labeling* can be seen by some as important tool for purposes of fostering identity, pride, solidarity, resistance, and reclamation.[18] Thus, another problem for a blanket policy to avoid usage of basic-level social terms is that it could take away an important vehicle for linguistic self-determination.

4.3 *Top-Down Adjustment*

Note that in all the cases of social labeling we've considered so far, there's nothing wrong with using words that pick out women, Koreans, Christians, and so on *per se*. What is problematic is rather the fact that the labels we use might indirectly communicate something inaccurate about the property distribution associated with the categories. Since the corresponding categories in fact *don't* have features that are common and distinct, the labels seem to make us susceptible to certain epistemic errors.

[18] For some literature on the topic, see Larkey *et al.* (1993); Yoder *et al.* (2011); Velazquez & Avila (2017); Galinsky *et al.* (2013); Boatswain & Lalonde (2000). See also Flores & Camp (2023) in this volume for discussion.

Consider an analogy. Cluster analysis is a common technique in statistics and data mining. When we perform cluster analysis, we aim, roughly, to group a set of objects in such a way that objects in the same group/cluster are more similar to each other than to those in other groups/clusters, where similarity or dissimilarity between data points is typically measured using a distance metric, such as Euclidean distance or cosine similarity. While there are different clustering methods (e.g., hierarchical clustering, k-means clustering), what they all have in common is to organize data based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Given the similarity to the Roschean model of categorization, it is unsurprising that clustering should, in principle, be susceptible to similar effects as labeling. Suppose the typical euclidean distance between clusters in your cluster analysis is d. Next, suppose someone tells you *that* some objects form a cluster c, but nothing else about that cluster. It would seem rational for you to expect that the distance between c and the nearest clusters to be roughly d. But when the *actual* distance is d^* such that $d^* << d$, the cluster-informant has led you epistemically astray.

So far, this situation seems roughly analogous to the situation we are in when it comes to social basic level terms. Luckily, it is unlikely that cluster analysis will have this kind of epistemic effect. That is, it'll hardly ever be the case that an analysis or analyst can only reveal to you *that* a set of objects falls into a cluster, but not much else about the properties of the cluster. At the same time, this imaginative scenario might uncover strategies to overcome basic level terms' potential to epistemically mislead. Suppose that sometimes, we are only able to communicate *that* a set of objects falls into a cluster. What could a theoretical statistician do to prevent recipients of this information to be led astray? Plausibly, they would want to make sure that the minimum distance between two clusters isn't substantially smaller than the typical distance other clusters have to each other. For example, if two clusters *are* too close to each other, you might want to merge them in order to make sure the minimal-distance requirement isn't violated.

Let's now return to labels and conceptual engineering. If we apply this idea to the case of labels and concepts, conceptual engineers would have to play the role of the theoretical statistician. In other words, in order to make sure the basic level terms in usage don't convey inaccurate information about property distributions, conceptual engineers would have to come up with *rules* about the minimal difference between categories that must exist in order to have basic-level status in our language. In addition, they might want to track which current labels in use follow the rules. That is, conceptual engineers would have to closely assess the correspondence between basic level terms in usage and actual statistical patterns in the world, in order to push for a reorganization of linguistic practices. For example, conceptual engineers could inform us which labels accurately reflect statistical similarity and distinctiveness structures, which ones don't, and introduce new basic level terms for social categories that would not lead to epistemic inaccuracies (akin to the merging of two clusters).

There are multiple problems with this proposal. For starters, it is easy to see that the idea is quite outlandish. Even if it would, in principle, fix the implementation challenge posed by social basic level terms, it would create implementation problems on its own. How feasible is it that some 'conceptual committee' can track feature distributions in the world and assess which people we should treat as groups, and which labels for them would be good? How feasible is it that some committee can devise top-down 'rules' about for when a term can be introduced into a community as basic-level? In addition, even if we could devise rules for when a label can be used in a community without the mentioned epistemic risks, most rules might lead to unwanted effects on its own. Take, again, our cluster-analysis analogy. In our hypothetical scenario, we considered merging different clusters if the euclidian distance to the closest cluster isn't sufficiently high. But this might mean that we create a cluster with little within-group similarity, compromising one desideratum for a 'good cluster'. And this, in turn, might lead to other epistemic problems: for example, we might assume that the cluster has less variance, and is more homogeneous, than it actually is.

As a final point, note that perceived properties of social groups *change*. Take gender categories. It is possible that, at various historical moments, the (perceived) difference and similarity between binary gender categories was significantly greater than it is now, and that ingroup variance and intergroup similarity have substantially increased at this point in time. But if properties of social categories change, a group that might be similar and distinct today might not be tomorrow. Needless to say, this would create additional feasibility problems for an ameliorative project that would try to track and make recommendations about proper use of social category labels in accordance with property distributions in the real world: our imaginary committee would constantly revise their recommendations for language use, depending on feature distributions in the real world.

4.4 Pessimism (+ Vigilance)

We've considered various options to remedy the effects of labels on our expectations regarding a categories' properties: expertise, label-eliminativism, and top-down adjustment. But we've also seen that all these proposals come with serious challenges. In light of the challenges, a final option emerges: embracing pessimism regarding the feasibility of ameliorating the effects of labels that are our focus. The preceding concerns suggest that we should simply accept that it is unavoidable that some of our representational devices will, at times, create misleading expectations about the property distribution of certain social groups. Among other things, as we have seen, whether or not labels introduce these can depend on all sorts of background conditions, including the actual property distribution associated with a group, or the contact between certain groups. Thus, instead of attempting to block *these* very effects of labels, conceptual engineers should rather accept them as a general limitation for ameliorative projects. In addition, they might redirect their focus on off-setting the bad effects that stem from labeling through countermeasures against other 'conceptual harms', such as the representation of groups on mainstream media.

Note that this 'pessimistic' strategy is only pessimistic about the possibility that a *general, systematic* solution can be found to block the inductive effects of basic level labels. But this still allows for the possibility that we push for certain ways of using certain labels on a case-by-case basis, citing as grounds the kind of effects this chapter was concerned with. For example, when noticing that not labeling a certain group as "woman" will have influences on the features we associate with members of that group, we can cite this as reason in favor of extending the application of the term to trans women. To give another example, in his book "Africa is not a country", Dipo Faloyin points out that many use terms such as "Africa" or "Africans" as *default* when describing people, peoples, or nations in Africa. This implies a wrong picture of Africa as a (religiously, culturally, economically, linguistically, etc.) homogeneous monolith (Faloyin 2022). Applying our framework, this usage conveys that the continent is where the sweet spot between similarity and distinctiveness lies. In both cases, it seems that greater terminological care and vigilance can contribute to mitigating certain stereotypes. But the cases also show that different ameliorative strategies are required for different situations, and that certain ameliorative techniques can lead to negative outcomes in other situations: as we learned earlier, *national* basic-level terms can give rise a distorted perception of property distributions too.

5 Closing Remarks and Future Avenues

The point of this paper was to contribute to the Feasibility Question in the field of conceptual engineering. Referring to social groups via basic-level terms, I argued, can systematically impact the way in which we represent those groups, by indirectly communicating something about the property structure of the group. This poses a challenge for ameliorative projects that aim to improve our category representations of social groups. I discussed various candidate approaches a conceptual engineer might advocate for in light of this challenge, and identified problems with each of them. The upshot is that conceptual engineers should pay close attention to the challenge and either find ways to overcome it, or, alternatively, accept that there is no one-size-fits-all solution to the challenge, and come up with case-by-case solutions to offset the epistemic costs of our conceptual practices.

In closing, I'd like to remark on a couple of important limitations of my preceding discussions, each of which I've so far ignored for the sake of simplification and development of my argument. First, a lot of argumentation in this chapter hinged on empirical work in the psychology of categorization. Much of the classic work in this area is from the second half of the twentieth century. I hope this chapter demonstrated the theoretical richness that lies in the work on categorization from this era, and served as a reminder of the importance to attend to the theoretical assumptions, insights, and motivations that constituted the foundation of the research program. At the same time, it is well-known that the psychological sciences have undergone many methodological advances in the past years. Thus, we should not only learn from and attend to the theoretical lessons in the history of psychology, but also keep revisiting these foundational questions in our theoretical *and empirical* work.

There's another limitation of this chapter. When introducing the hierarchy of concepts and the notion of *basic level* concepts and terms, we've only considered concepts from animal, plant, and inanimate object domains. As a matter of fact, Rosch *et al.* (1976)'s important studies that uncovered the so-called 'basic level advantage' only studied non-social categories. Nevertheless, much of this chapter simply took for granted that the framework can be extended and applied to social categories. It is unclear to which extent this move is warranted. Evidence about the extent to which social categories are organized hierarchically is mixed.[19] On the one hand, theorists have pointed out that the Principles should clearly extend to the social domain. As Murphy & Lassaline (1997) put it,

> Person classification serves many of the same functions as does object categorization. As with objects, information can be reduced to a manageable level by using only some characteristic to classify a membership in a particular category. [...] Person classification, however, also has the same cost as object categorization. By treating nonequivalent things as equivalent, information about individuals is lost. (Murphy & Lassaline 1997, p. 117)[20]

Correspondingly, it seems easy to think of a 'medium' level of specificity when it comes to social categories (e.g., "teacher" seems more natural than "educator" (superordinate) or "kindergarten teacher" (subordinate)). In addition, some studies present direct evidence that social categories can be organized in different levels of specificity.[21] For example, focusing on gender categories, several studies documented evidence that woman functions as a basic-level category with rich category associations (Deaux *et al.* 1985; Harper & Schoeman 2003).[22]

[19] For an old, but insightful, discussion on whether models of natural object concepts extend to social concepts, see Lingle *et al.* (1984).

[20] See also Leslie (2017): "[...] we can generalize the idea of a basic-level kind to the social arena: these will be social kinds that are perceived to have essences that occupy a "sweet spot" in trade-offs between distinctiveness (which is compromised as groups become less inclusive) and predictiveness (in the sense of grounding the maximal number of common features—a feature that is compromised as groups become more inclusive). [...] Such social kinds, we may suppose, will surely include racial, ethnic, and religious groups" (Leslie 2017, pp. 409–410).

[21] For example, Cantor & Mischel (1979) reported that Rosch's hierarchy can indeed be extended to social categories: in feature listing tasks, there was a middle level that had many more features listed for a category than the superordinate level, and a subordinate level with only few additional features. Similarly, Cantor *et al.* (1980) found that clinicians' psychodiagnostic categories can be organized in a hierarchical way, with a distinctive basic level the categories of which are both informative and distinctive. There's evidence that trait concepts (John *et al.* 1991) and emotion concepts (Shaver *et al.* 1987; Bretherton & Beeghly 1982) also exhibit Rosch's hierarchy effect, with a basic level that is privileged (e.g., in the case of emotions, basic emotions such as *anger, joy, fear* were at an intermediate, basic level.

[22] In fact, Deaux *et al.* (1985) tested the hypothesis that gender categories such as woman and man are *not* a basic-level category, on the grounds that they seem like categories so broad that it's unlikely for them to occupy a middle-level of inclusivity. Contrary to their hypothesis (but, in my view, still unsurprisingly), however, they found that these gender categories are associated with rich category associations, which speaks in favor of the basic-level hypothesis.

At the same time, various theorists have emphasized that social concepts can't be as neatly organized into a hierarchical taxonomy as, say, animal or plant concepts (Lingle *et al.* 1984; Holyoak & Gordon 1984; Murphy & Lassaline 1997). Many non-social categories are non-overlapping and mutually exclusive across the same horizontal axis: If you are a lion, you aren't a dog. Social concepts, in contrast, exhibit greater flexibility and purpose-dependence, and are intersective even when they inhabit the same level of specificity. If you're a teacher, you can also be a man, and an Estonian, and a father, and a dancer (Lingle *et al.* 1984; Murphy & Lassaline 1997; Fiske & Taylor 2013). While this doesn't mean that for *each* of these attributes, there isn't a medium level of specificity that's used as default and has the potential to distort our perception of property distributions, we need more work (much empirical work on the question is pre-2000!) that investigates the relationship between basic level concepts, terms, and social categories.

With this in mind, my chapter can be understood as an exercise in speculative psychology. I extended insights from foundational work in the psychology of concepts to the social domain, and extracted challenges for conceptual engineers who aim to ameliorate these very concepts. The general lesson of my paper, and the imperative for conceptual engineers and adjacent areas, stands. More work is needed to uncover the effects of social basic level terms on our perceptions of property distributions (and whether they align with the predictions made in this paper), and to meet the challenges that grow out of these.

Acknowledgments

I thank Guillermo Del Pinal, Joshua Knobe, Laura Schroeter, François Schroeter, Karen Jones, and the audience at the *ICCE2* online workshop at St. Andrews-Arché for useful feedback and comments on earlier versions and presentations of this paper. Discussions I've had at the *Conceptual Engineering Online Workshop* at the University of Oslo and during my 2021 upper-division seminar *Stereotypes and Bias in Minds and Machines* at UIUC have given rise to numerous ideas developed in this paper.

Works Cited

Allport, Gordon W. 1954. *The nature of prejudice*. Addison-Wesley.

Barnes, Elizabeth. 2016. *The minority body: A theory of disability*. Oxford University Press.

Beeghly, Erin. 2015. What is a stereotype? what is stereotyping? *Hypatia*, **30**(4), 675–691.

Berlin, Brent. 2014. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Vol. 185. Princeton University Press.

Bertrand, Marianne, & Duflo, Esther. 2017. Field experiments on discrimination. *Handbook of economic field experiments*, **1**, 309–393.

Boatswain, Sharon J, & Lalonde, Richard N. 2000. Social identity and preferred ethnic/racial labels for blacks in canada. *Journal of black psychology*, **26**(2), 216–234.

Bordalo, Pedro, Coffman, Katherine, Gennaioli, Nicola, & Shleifer, Andrei. 2016. Stereotypes. *The quarterly journal of economics*, **131**(4), 1753–1794.

Borges, Jorge Luis. 1937. The analytical language of john wilkins. *Other inquisitions*, **1952**, 101–105.

Bosse, Anne. 2022. Stereotyping and generics. *Inquiry*, 1–17.

Bretherton, Inge, & Beeghly, Marjorie. 1982. Talking about internal states: The acquisition of an explicit theory of mind. *Developmental psychology*, **18**(6), 906.

Brown, Roger. 1958. How shall a thing be called? *Psychological review*, **65**(1), 14.

Burge, Tyler. 1993. Concepts, definitions, and meaning. *Metaphilosophy*, **24**(4), 309–325.

Byrne, Alex. 2020. Are women adult human females? *Philosophical studies*, **177**(12), 3783–3803.

Callanan, Maureen A. 1985. How parents label objects for young children: The role of input in the acquisition of category hierarchies. *Child development*, 508–523.

Cantor, Nancy, & Mischel, Walter. 1979. Prototypes in person perception. *Pages 3–52 of: Advances in experimental social psychology*, vol. 12. Elsevier.

Cantor, Nancy, Smith, Edward E, French, Rita D, & Mezzich, Juan. 1980. Psychiatric diagnosis as prototype categorization. *Journal of abnormal psychology*, **89**(2), 181.

Carnaghi, Andrea, Maass, Anne, Gresta, Sara, Bianchi, Mauro, Cadinu, Mara, & Arcuri, Luciano. 2008. Nomina sunt omina: on the inductive potential of nouns and adjectives in person perception. *Journal of personality and social psychology*, **94**(5), 839.

Catapang Podosky, Paul-Mikhail. 2022. Can conceptual engineering actually promote social justice? *Synthese*, **200**(2), 160.

Clarke, Alex, & Tyler, Lorraine K. 2015. Understanding what we see: How we derive meaning from vision. *Trends in cognitive sciences*, **19**(11), 677–687.

Deaux, Kay, Winton, Ward, Crowley, Maureen, & Lewis, Laurie L. 1985. Level of categorization and content of gender stereotypes. *Social cognition*, **3**(2), 145–167.

Del Pinal, Guillermo, Madva, Alex, & Reuter, Kevin. 2017. Stereotypes, conceptual centrality and gender bias: An empirical investigation. *Ratio*, **30**(4), 384–410.

Dembroff, Robin A. 2016. What is sexual orientation? *Philosophers' imprint*, **16**.

Devos, Thierry, Comby, Loraine, & Deschamps, Jean-Claude. 1996. Asymmetries in judgements of ingroup and outgroup variability. *European review of social psychology*, **7**(1), 95–144.

Dixon, John, Durrheim, Kevin, & Tredoux, Colin. 2005. Beyond the optimal contact strategy: A reality check for the contact hypothesis. *American psychologist*, **60**(7), 697.

Dovidio, John F, Gaertner, Samuel L, & Kawakami, Kerry. 2003. Intergroup contact: The past, present, and the future. *Group processes & intergroup relations*, **6**(1), 5–21.

Downing, Pamela A. 1977. On" basic levels" and the categorization of objects in english discourse. *Pages 475–487 of: Annual meeting of the berkeley linguistics society*, vol. 3.

Faloyin, Dipo. 2022. *Africa is not a country: Notes on a bright continent*. W. W. Norton & Company.

Fischer, Eugen. 2020. Conceptual control: On the feasibility of conceptual engineering. *Inquiry: An interdisciplinary journal of philosophy*, 1–29.

Fiske, Susan T, & Taylor, Shelley E. 2013. *Social cognition: From brains to culture*. Sage.

Flores, Carolina, & Camp, Elisabeth. 2023. "that's all you really are": Centering identities without essentialist beliefs. *In:* Haslanger, Sally, Jones, Karen, Schroeter, Francois, & Schroeter, Laura (eds), *Mind, language, and social hierarchy*. Oxford University Press.

Galinsky, Adam D, Wang, Cynthia S, Whitson, Jennifer A, Anicich, Eric M, Hugenberg, Kurt, & Bodenhausen, Galen V. 2013. The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling. *Psychological science*, **24**(10), 2020–2029.

Gelman, Susan A, & Heyman, Gail D. 1999. Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological science*, **10**(6), 489–493.

Gennaioli, Nicola, & Shleifer, Andrei. 2010. What comes to mind. *The quarterly journal of economics*, **125**(4), 1399–1433.

Gilbert, Gustave M. 1951. Stereotype persistence and change among college students. *The journal of abnormal and social psychology*, **46**(2), 245.

Gosselin, Frédéric, & Schyns, Philippe G. 2001. Why do we slip to the basic level? computational constraints and their implementation. *Psychological review*, **108**(4), 735.

Hampton, James A. 1982. A demonstration of intransitivity in natural categories. *Cognition*, **12**(2), 151–164.

Harper, Marcel, & Schoeman, Wilhelm J. 2003. Influences of gender as a basic-level category in person perception on the gender belief system. *Sex roles*, **49**, 517–526.

Haslanger, Sally. 2000. Gender and race: (what) are they? (what) do we want them to be? *Noûs*, **34**(1), 31–55.

Holyoak, Keith J., & Gordon, Peter C. 1984. Information processing and social cognition. *Pages 39–70 of:* Wyer, Robert S., Jr., & Srull, Thomas K. (eds), *Handbook of social cognition*, vol. 1. Lawrence Erlbaum Associates Publishers.

Hopkins, Daniel J. 2010. Politicized places: Explaining where and when immigrants provoke local opposition. *American political science review*, **104**(1), 40–60.

Horton, M. S., & Markman, E. M. 1980. Developmental differences in the acquisition of basic and superordinate categories. *Child development*, **51**, 708–719.

Isaac, Manuel Gustavo. 2020. How to conceptually engineer conceptual engineering? *Inquiry*, 1–24.

Isaac, Manuel Gustavo, Koch, Steffen, & Nefdt, Ryan. 2022. Conceptual engineering: A road map to practice. *Philosophy compass*, **n/a**(n/a), e12879.

John, Oliver P, Hampson, Sarah E, & Goldberg, Lewis R. 1991. The basic level in personality-trait hierarchies: studies of trait use and accessibility in different contexts. *Journal of personality and social psychology*, **60**(3), 348.

Johnson, Gabbrielle M. 2020. The structure of bias. *Mind*, **129**(516), 1193–1236.

Johnson, Kathy E, & Mervis, Carolyn B. 1997. Effects of varying levels of expertise on the basic level of categorization. *Journal of experimental psychology: General*, **126**(3), 248.

Johnston, Mark, & Leslie, Sarah-Jane. 2012. Concepts, analysis, generics and the canberra plan. *Philosophical perspectives*, **26**(1), 113–171.

Johnston, Mark, & Leslie, Sarah-Jane. 2019. 7 cognitive psychology and the metaphysics of meaning. *Metaphysics and cognitive science*.

Jolicoeur, Pierre, Gluck, Mark A, & Kosslyn, Stephen M. 1984. Pictures and names: Making the connection. *Cognitive psychology*, **16**(2), 243–275.

Jones, Gregory V. 1983. Identifying basic categories. *Psychological bulletin*, **94**(3), 423.

Karlins, Marvin, Coffman, Thomas L, & Walters, Gary. 1969. On the fading of social stereotypes: studies in three generations of college students. *Journal of personality and social psychology*, **13**(1), 1.

Katz, Daniel, & Braly, Kenneth. 1933. Racial stereotypes of one hundred college students. *The journal of abnormal and social psychology*, **28**(3), 280.

Kenworthy, Jared B, Turner, Rhiannon N, Hewstone, Miles, & Voci, Alberto. 2005. Intergroup contact: When does it work, and why. *On the nature of prejudice: Fifty years after allport*, 278–292.

Koch, Steffen. 2021. Engineering what? on concepts in conceptual engineering. *Synthese*, **199**(1), 1955–1975.

Koslow, Allison. 2022. Meaning change and changing meaning. *Synthese*, **200**(2), 1–26.

Larkey, Linda Kathryn, Hecht, Michael L, & Martin, Judith. 1993. What's in a name? african american ethnic identity terms and self-determination. *Journal of language and social psychology*, **12**(4), 302–317.

Leslie, Sarah-Jane. 2017. The original sin of cognition: Fear, prejudice, and generalization. *The journal of philosophy*, **114**(8), 393–421.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Lingle, J. H., Altom, M. W., & Medin, D. L. 1984. Of cabbages and kings: Assessing the extendibility of natural object concept models to social things. *Pages 71–117 of:* Wyer, R. S., Jr., & Srull, T. K. (eds), *Handbook of social cognition*, vol. 1. Lawrence Erlbaum Associates Publishers.

Linville, Patricia W, Fischer, Gregory W, & Salovey, Peter. 1989. Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of personality and social psychology*, **57**(2), 165.

Löhr, Guido. 2020. Concepts and categorization: do philosophers and psychologists theorize about different things? *Synthese*, **197**(5), 2171–2191.

Machery, Edouard. 2009. *Doing without concepts*. Oxford University Press.

Machery, Edouard. 2017. *Philosophy within its proper bounds*. Oxford University Press.

Machery, Edouard. 2021. A new challenge to conceptual engineering. *Inquiry*, 1–24.

Madon, Stephanie, Guyll, Max, Aboufadel, Kathy, Montiel, Eulices, Smith, Alison, Palumbo, Polly, & Jussim, Lee. 2001. Ethnic and national stereotypes: The princeton trilogy revisited and revised. *Personality and social psychology bulletin*, **27**(8), 996–1010.

Madva, Alex, & Brownstein, Michael. 2018. Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs*, **52**(3), 611–644.

Manne, Kate. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Markman, Arthur B, & Wisniewski, Edward J. 1997. Similar and different: The differentiation of basic-level categories. *Journal of experimental psychology: Learning, memory, and cognition*, **23**(1), 54.

McKeown, Shelley, & Dixon, John. 2017. The "contact hypothesis": Critical reflections and future directions. *Social and personality psychology compass*, **11**(1), e12295.

Medin, Douglas L, Lynch, Elizabeth B, & Solomon, Karen O. 2000. Are there kinds of concepts? *Annual review of psychology*, **51**(1), 121–147.

Mervis, Carolyn B, & Crisafi, Maria A. 1982. Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child development*, 258–266.

Mervis, Carolyn B, & Rosch, Eleanor. 1981. Categorization of natural objects. *Annual review of psychology*, **32**(1), 89–115.

Messick, David M, & Mackie, Diane M. 1989. Intergroup relations. *A psychology*, **40**, 45–81.

Mullen, Brian, & Hu, Li-Tze. 1989. Perceptions of ingroup and outgroup variability: A meta-analytic integration. *Basic and applied social psychology*, **10**(3), 233–252.

Murphy, Gregory. 2004. *The big book of concepts*. MIT press.

Murphy, Gregory. 2023. Categories and concepts. *In:* Biswas-Diener, Robert, & Diener, Ed (eds), *Noba textbook series: Psychology*. Champaign, IL: DEF publishers.

Murphy, Gregory L, & Brownell, Hiram H. 1985. Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of experimental psychology: Learning, memory, and cognition*, **11**(1), 70.

Murphy, Gregory L, & Lassaline, Mary E. 1997. Hierarchical structure in concepts and the basic level of categorization. *Knowledge, concepts, and categories*, 93–131.

Murphy, Gregory L, & Smith, Edward E. 1982. Basic-level superiority in picture categorization. *Journal of verbal learning and verbal behavior*, **21**(1), 1–20.

Nefdt, Ryan M. 2021. Concepts and conceptual engineering: answering cappelen's challenge. *Inquiry*, 1–29.

Neufeld, Eleonore. 2019. *An essentialist theory of the meaning of slurs*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Neufeld, Eleonore. 2020a. Can we perceive mental states? *Synthese*, **197**(5), 2245–2269.

Neufeld, Eleonore. 2020b. Pornography and dehumanization: The essentialist dimension. *Australasian journal of philosophy*, **98**(4), 703–717.

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

Neufeld, Eleonore. 2023. Engineering social concepts: Feasibility and causal models. *ms*.

Neufeld, Eleonore, Brown, Elliot C, Lee-Grimm, Sie-In, Newen, Albert, & Brüne, Martin. 2016. Intentional action processing results from automatic bottom-up attention: An eeg-investigation into the social relevance hypothesis using hypnosis. *Consciousness and cognition*, **42**, 101–112.

Pettigrew, Thomas F, & Tropp, Linda R. 2006. A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, **90**(5), 751.

Pettigrew, Thomas F, Tropp, Linda R, Wagner, Ulrich, & Christ, Oliver. 2011. Recent advances in intergroup contact theory. *International journal of intercultural relations*, **35**(3), 271–280.

Pinder, Mark. 2022. Is haslanger?s ameliorative project a successful conceptual engineering project? *Synthese*, **200**(4), 1–22.

Plunkett, David, & Cappelen, Herman. 2020. A guided tour of conceptual engineering and conceptual ethics. *Pages 1–26 of:* Cappelen, Herman, Plunkett, David, & Burgess, Alexis (eds), *Conceptual engineering and conceptual ethics*. Oxford: Oxford University Press.

Puddifoot, Katherine. 2021. *How stereotypes deceive us*. Oxford University Press.

Riggs, Jared. 2019. Conceptual engineers shouldn't worry about semantic externalism. *Inquiry*, 1–22.

Ritchie, Katherine. 2021. Essentializing language and the prospects for ameliorative projects. *Ethics*, **131**(3), 460–488.

Rosch, Eleanor. 1978. Principles of categorization. *Pages 28–48 of:* Rosch, E., & Lloyd, B. B. (eds), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Rosch, Eleanor, Mervis, Carolyn B, Gray, Wayne D, Johnson, David M, & Boyes-Braem, Penny. 1976. Basic objects in natural categories. *Cognitive psychology*, **8**(3), 382–439.

Saul, Jennifer. 2017. Are generics especially pernicious? *Inquiry*, 1–18.

Shaver, Phillip, Schwartz, Judith, Kirson, Donald, & O'connor, Cary. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, **52**(6), 1061.

Smith, Madorah Elizabeth. 1926. *Investigation of the development of the sentence and the extent of vocabulary in young children*. Vol. 3. The University.

Stock, Kathleen. 2018. Changing the concept of "woman" will cause unintended harms. *The economist*, **6**.

Tanaka, James W, & Taylor, Marjorie. 1991. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, **23**(3), 457–482.

Tversky, Amos. 1977. Features of similarity. *Psychological review*, **84**(4), 327.

Tversky, Amos, & Kahneman, Daniel. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, **90**(4), 293.

Velazquez, Efren, & Avila, Melissa. 2017. Ethnic labels, pride, and challenges: A qualitative study of latinx youth living in a new latinx destination community. *Journal of ethnic and cultural studies*, **4**(1), 1.

Wang, Xiaoxi, Ma, Xie, Tao, Yun, Tao, Yachen, & Li, Hong. 2018. How semantic radicals in chinese characters facilitate hierarchical category-based induction. *Scientific reports*, **8**(1), 5577.

Westra, Evan. 2019. Stereotypes, theory of mind, and the action?prediction hierarchy. *Synthese*, **196**(7), 2821–2846.

Whiteley, Ella. 2023. A woman first and a philosopher second: Relative attentional surplus on the wrong property. *Ethics*.

Wisniewski, Edward J, & Murphy, Gregory L. 1989. Superordinate and basic category names in discourse: A textual analysis. *Discourse processes*, **12**(2), 245–261.

Yoder, Janice D, Tobias, Ann, & Snell, Andrea F. 2011. When declaring "i am a feminist" matters: Labeling is linked to activism. *Sex roles*, **64**, 9–18.

Zingora, Tibor, Vezzali, Loris, & Graf, Sylvie. 2021. Stereotypes in the face of reality: Intergroup contact inconsistent with group stereotypes changes attitudes more than stereotype-consistent contact. *Group processes & intergroup relations*, **24**(8), 1284–1305.

Zubin, David, & Köpcke, Klaus-Michael. 1986. Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. *Noun classes and categorization*, **139**, 180.