Engineering Social Concepts: Feasibility and Causal Models

Eleonore Neufeld

University of Massachusetts Amherst

February 29, 2024

Abstract

How feasible are conceptual engineering projects of social concepts that aim for the engineered concept to be widely adopted in ordinary everyday life? Predominant frameworks on the psychology of concepts that shape work on stereotyping, bias, and machine learning have grim implications for the prospects of conceptual engineers: conceptual engineering efforts are ineffective in promoting certain social-conceptual changes. Specifically, since conceptual components that give rise to problematic social stereotypes are sensitive to statistical structures of the environment, purely conceptual change won't be possible without corresponding world change. This tradition, however, tends to ignore that concepts don't only encode statistical, but also causal information. Paying attention to this feature of concepts, I argue, shows that conceptual engineering is not only possible. There is an imperative to conceptually-engineer.

1 Introduction

Academic practice is full of examples of conceptual engineering—i.e., proposals for how to change our representational devices, such as words and concepts, for the better. For example, Chomsky has famously argued that within linguistic-scientific contexts, we should treat 'language' as describing an internal, intensional, and individual property of human psychology (Chomsky 1980; Chomsky *et al.* 2000). Importantly, Chomsky's aim wasn't to merely delineate the phenomenon he was interested in. Within scientific contexts, he insists, 'i-language' is the only subject

matter that our inquiry can—*ought*—target. His proposal hasn't been without success. Arguably, this meaning of "language" is currently dominating academic linguistics.

Some conceptual engineering efforts target *theoretical* vocabulary in relatively confined communities, such as academic fields. Chomsky's engineering of "language" in linguistics is one such example. And there are many others. Sally Haslanger's 2000 proposal to change "woman" or Robin Dembroff's 2016 proposal to change "sexual orientation", are both in the service, the authors claim, of improving *theorizing* in socio-political theory. Because the target populations (e.g., academic fields) are fairly confined in all these cases, it is relatively straightforward to see how successful conceptual change could be accomplished, at least if the proposals are met with a sufficient degree of acceptance.[1]

Other conceptual engineering projects don't aim at merely changing representations employed in theoretical discourse. Instead, they aspire to change concepts of 'the folk'—ordinary concepts we use in our day-to-day life to guide our categorization and induction behavior. There's no shortage of examples for such efforts. WOMAN, MARRIAGE, IMMIGRANT, FOOD, DISABILITY, and more have all been submitted as candidates for improvement by conceptual activists inside and outside philosophy. But changing ordinary concepts of the folk is an entirely different enterprise compared to changing theoretical vocabulary used in more 'artificial' settings. How is it possible to enact these changes in our ordinary conceptual practices, given that we are already bound to this or that conceptual practice? This problem is widely-recognized within the literature on conceptual engineering as the 'feasibility question' (Fischer 2020; Machery 2021):[2]

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

1. This is, of course, not always the case. For example, Haslanger's proposal for "woman" has sparked an extensive debate, including many revisions and counter-proposals to Haslanger's ameliorative analysis (see, e.g., Jenkins 2016; Díaz León 2019).

2. It is important to distinguish the feasibility question from what is sometimes called the *implementation challenge* (or implementation problem) for conceptual engineering (Jorem 2021; Cappelen 2018). While feasibility questions are about "the possibility of modifying concepts in light of *contingent facts about psychology and the social world*" (Machery 2021, p. 7, emphasis added) and thus informed by empirical facts about the social world or psychology, the implementation problem worries about the *abstract* possibility of conceptual engineering in light of certain metasemantic (viz., externalist) views about meaning and meaning-change. See Riggs (2019); Burgess & Plunkett (2013); Koch (2021b); Cappelen (2018); Jorem (2021); Deutsch (2020) for discussions of the implementation problem.

As Machery (2021) points out, addressing the Feasibility Question is an endeavour in non-ideal theorizing, since it takes as starting point contingent facts about the actual world. For the same reason, the Feasibility Question is of special relevance for *social* categories. Since our induction and categorization practices can have direct material and normative consequences for members of the relevant social groups, the question of whether, and if so how, we can implement changes in these practices has obvious urgency.

The purpose of this paper is to shed light on precisely this question: How feasible are conceptual engineering projects of social concepts that aim for the engineered concept to be adopted in ordinary everyday life? In response to this question, I show that a major tradition in the science of categorization implies that conceptual engineering efforts are *ineffective* in promoting conceptual change (§2 and §3). Next, I show that although this tradition has a huge influence on current theorizing about social concepts, it tends to ignore that concepts don't only encode statistical, but also causal information (§4 and §5). If we pay attention to this feature of concepts, I argue in §6, we are left with a somewhat surprising result. Conceptual engineering is not only possible. Instead, there is an imperative to conceptually-engineer.

Before we start, a few important clarifications are in order. First, in this paper, I operate with a 'practical-aim' approach to conceptual engineering, according to which "[c]onceptual engineers [...] aim to change how people think about objects, how they classify them, and how they use words (e.g. by getting people to stop calling whales 'fish', or start calling certain acts 'misogynistic')" (Koch 2021a, p. 1958). Second, in my view, the practical aim approach pairs best with a psychological approach to conceptual engineering, according to which "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 2021a, p. 1956). The psychological practical-aim approach to conceptual engineering has been adopted, defended, and developed by various philosophers in the literature on conceptual engineering (Isaac 2020, 2021b,a; Isaac *et al.* 2022; Fischer 2020; Kitsik forthcoming; Machery 2017, 2021; Quilty-Dunn 2021). Thirdly, in line with the psychological practical-aim approach to conceptual engineering, I use "concepts" to describe indiviual-level psychological entities; specifically, bodies

of information about x that are stored in long term memory and retrieved *by default*[3] in processes underlying most, if not all, higher cognitive competencies (e.g., inductive reasoning and categorization) when these processes result in judgements about x.[4] This characterization corresponds to the favored use in the psychological tradition of conceptual engineering, and is continuous with the dominant use in the psychological literature on concepts (cf. Murphy 2004; Johnston & Leslie 2019).

It is important to emphasize that the psychological approach to conceptual engineering doesn't imply that any sort of belief change will amount to conceptual engineering.[5] In fact, this approach gives us a straightforward way to distinguish conceptual engineering from mere belief change. Suppose I learn that there are 12.4 billion tables on the planet. This would lead me to form a new belief about tables. Nevertheless, this wouldn't suffice to make this piece of information retrieved *by default* when categorizing tables, reasoning about tables, making inductive inferences about tables, and so on. Thus, merely acquiring the belief that there are 12.4 billion tables on the planet would not amount to conceptually engineering TABLE. In contrast, information such as +HAS LEGS or +HAS A TOP *is* retrieved by default in cognitive activities involving TABLE. Hence, change in the latter kind of information would amount to conceptual engineering.[6]

3. By "by default", I mean that they are retrieved spontaneously, automatically, quickly, and systematically (cf. Machery 2015). Note that under the present operationalization of 'conceptual engineering', changes in, e.g., people's stereotypes or implicit biases about social groups would amount to conceptual engineering.

4. This characterization leans heavily on the one developed and defended in Edouard Machery's work (cf. Machery 2009, 2015, 2017, 2021, 2022).

5. Relatedly, note also that conceptual engineers who adopt a psychological approach *don't need to be* committed to a molecular view in the debate on the structure of concepts within the philosophy of cognitive science. This is simply because in that debate, "concept" is not used in the technical sense employed in the literature on psychological conceptual engineering outlined earlier. For this reason, the psychological approach to conceptual engineering is, as such, not subject to well-known objections against molecular views of concepts, such as arguments from compositionality, disagreement, and communicative success. Those conceptual engineers who *are* independently committed to molecular views will, of course, have to face these objections and/or make use of available arguments that have been offered in response to these concerns (e.g., Del Pinal 2016, 2018; Jönsson 2017; Kamp & Partee 1995).

6. That said, readers who have strong views about the proper domain of conceptual engineering and think the psychological approach doesn't amount to 'real' conceptual engineering (e.g., Cappelen 2020) can simply read this paper as investigating the question of whether, and to which extent, certain changes in the bodies of information that are intimately tied to certain concepts, are stubborn, retrieved fast, and automatically, and have pervasive downstream effects on cognition and behavior, are possible.

2 Background: Psychology of Concepts

In order to find out how we can change the "psychological structures that explain our mental and linguistic behavior" (Koch 2021a), we have to know what these structures look like. The psychological tradition that responds to this question is deeply rooted in Eleanor Rosch's pioneering work on categorization (Mervis & Rosch 1981; Rosch 1978; Murphy 2004; Murphy & Lassaline 1997). The core insight of Rosch's work was that human categorization is not arbitrary. Instead, at least two fundamental psychological principles constrain possible systems of classification for all human cultures. As we will see, these principles have crucial implications for our theories of categorization and, *inter alia*, the feasibility of conceptual engineering projects. The principles are the *Principle of Cognitive Economy* and the *Principle of Perceived World Structure* (Rosch 1978):

1. **Principle of Cognitive Economy.** The task of conceptual systems is to provide maximum information with the least cognitive effort.

2. **Principle of Perceived World Structure.** The perceived world comes as structured information rather than as arbitrary and unpredictable attributes.

Let us unpack each principle in turn.

The first principle strikes a compromise between two distinct pressures our cognitive systems are subject to: to get as much information as possible from an act of categorization, and to preserve finite cognitive resources. From this principle, it follows 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' 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.

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 highly similar to each other. 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 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 infer it's a dog; once you detect that something has a very long neck, you can infer it's a giraffe.

Let's now turn to the Principle of Perceived World Structure. 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.

Following Rosch (1978), we can use Jorge Luis Borges' fictional animal taxonomy *Celestial Emporium of Benevolent Knowledge* to illustrate the wide explanatory reach of the Principles (Borges 1937):

> the animals are divided into: (a) belonging to the emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies.

What's remarkable about this taxonomy is that it does not exist. No human culture has lexical concepts that pick out the categories listed in Borges' taxonomy. The Principles of Categorization explain why. 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 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 so on.

Within philosophy, the Principles are closely aligned with the empiricist tradition, according to which our concepts mirror statistical regularities of environmental input. Within psychology, the two Principles are intimately connected to the well-known Prototype Theory of concepts, according to which we represent categories in terms of typical features. Importantly, typicality is simply another way of saying that features have an optimal degree of across-category distinctiveness and within-category similarity. Correspondingly, in the rest of the paper, I'll use "typical" to refer to conceptual features that are high in similarity and/or distinctiveness.

3 Implications for Conceptual Engineering

What are the implications of the two Principles for the prospects of conceptual engineering? Again, it is best to illustrate this via an example.

Consider the concept pitbull.[7] If we have an empiricist view of social concepts as suggested by the Principles, and view them as constituted by or associated with simple prototypes, the prototype of pitbull should simply mirror the co-occurrences of the perceived world. Given the real-world input we receive through various sources (e.g., perceptual real-world input, verbal testimony and pictorial representations via various media outlets), our concept might be associated with typical features such as +short coat, +muscular, +aggressive, and so on. Suppose that our present conceptual engineering aim is to change some of these typical features. In particular, we want to change the concept such that it doesn't encode +aggressive any more, but is instead associated with +friendly. This means that the new feature would strongly shape people's ordinary induction behavior. Instead of automatically inferring that a pitbull is dangerous, they will be disposed to infer that a pitbull is friendly. Suppose further that we succeed in implementing such a change. Will this mean we succeeded in our conceptual engineering efforts?

No, for the following reason. *If the perceived world pattern stays unchanged*, then, given the Principles, the new concept will quickly revert back to the old one. If various input sources represent pitbulls as dangerous, this will be reflected in the features associated with the concept and their statistical weights. This upshot, it seems, has important consequences for the prospects of conceptual engineering.

7. Even if pitbull is arguably not a straightforward example of a social concept, it is easy to see that the points I illustrate via this example directly extend to paradigmatic social concepts.

Novel ordinary concepts (i.e., classification and induction proposals) won't be implemented if the real-world input we receive aligns better with old concepts. Thus, if our ameliorative aim is to change a given conceptual practice, such a change will fail to stick if the new classification system is not exposed to the corresponding data bundles. Conceptual engineering projects for ordinary social concepts would be at an impasse.

Importantly, the Principles of Categorization are extremely influential in areas of psychology in which our understanding of social category representations is essential—including developmental, social, and cognitive psychology. The principles loom large in dominant theories of explicit and implicit bias, stereotyping, and even algorithmic bias. For example, in his *TED* talk "Can Prejudice Ever Be a Good Thing?" (viewed more than 1.5 million times), Paul Bloom (2014) contends that

> Our ability to stereotype people is not some sort of arbitrary quirk of the mind, but rather it's a specific instance of a more general process, which is that we have experience with things and people in the world that fall into categories, and we can use our experience to make generalizations about novel instances of these categories. So everyone here has a lot of experience with chairs and apples and dogs, and based on this, you could see these unfamiliar examples and you could guess — you could sit on the chair, you could eat the apple, the dog will bark.

Similarly, in their famous book *Blindspot*, Banaji & Greenwald (2013) provide the following analysis of the workings of the *Implicit Association Test* ('IAT'):[8]

> [The IAT's] effectiveness relies on the fact that your brain has stored years of past experiences that you cannot set aside when you do the IAT's sorting tasks. (Banaji & Greenwald 2013, p. 66)

And, the Principles also influenced work that aims to draw insights about human social categorization tendencies from the study of machine bias. For example, in a recent paper by Johnson (2021), she argues that human bias, just like machine

8. The Implicit Association Test (or IAT, for short) is a psychological reaction-time measure in which subjects are instructed to sort words or images into categories. Differences in error-rates or speed for stereotype-consistent and stereotype-inconsistent trials are taken to reveal differences in strength of associations between categories and attributes. See Holroyd *et al.* (2017); Nosek *et al.* (2011); Brownstein *et al.* (2019); Banaji & Greenwald (2013) for examples, overviews, and discussions.

bias, is a result of the training data on the basis of which we make predictions, following the slogan "garbage in, garbage out" (p. 9948). Thus, without a change of input data (i.e., social structures), our biases—i.e., induction and categorization practices—will remain unchanged. According to Johnson, this "can serve to bolster a critical insight from equality advocates that *representation matters*" (p. 9956, emphasis orig.).

Given that dominant theories of social category representations in psychology are so intertwined with the Principles, it is unsurprising that stereotype *intervention* strategies build on them, too. Cognitive intervention strategies focus on providing counterstereotypic or nonstereotypic information about group members to undermine or dilute stereotypic association.[9] The idea is that while a lot of your perceived world patterns has consisted of <PITBULL, AGGRESSIVE> pairings, we now provide you with input data that consist of <PITBULL, FRIENDLY> pairings.

All this, we might think, leads to a fairly grim conclusion for psychology-focused conceptual engineering projects that aim for better ordinary category representations of social groups. Namely, that conceptual change, including change in our conceptual practices that relate to stereotyping and bias, can only be implemented via change in the perceived world. Thus, there is no hope that we can *engineer* better ordinary social concepts, conditional on world patterns being fixed. In other words: there is no way of *directly* manipulating conceptual content, independently of what the world is like. Instead, the only way to change conceptual structures is by changing input data bundles. Conceptual engineers, therefore, are out of a job. Instead of being invested in distinctively *conceptual* activism, they are better served by focusing on changing structures and systems in the real world.[10]

In response to this concern, advocates of psychological approaches to conceptual engineering might stress that this conclusion is too hasty, and point to recent proposals of how we can overcome the challenge of changing problematic conceptual features that are rooted in statistical real-world patterns. Specifically, Fischer (2020) has recently argued that, in contexts of language comprehension, changes to the *linguistic environment* in which a to-be engineered lexical item is embedded can inhibit or enrich stereotypical information retrieved when linguistically processing

9. See Dovidio *et al.* (2000); Lai *et al.* (2016); Gonzalez *et al.* (2021); Rothbart & John (1985); Weber & Crocker (1983).

10. See Neufeld (forthcoming) for further discussion of other consequences for conceptual engineering projects that follow from Rosch's Principles.

that lexical item in a way that's consistent with ameliorative aims for that item (see also Isaac 2021a,b; Fischer & Engelhardt 2017). When a linguistic context is incompatible with a certain aspect of the retrieved stereotype, that aspect will be *inhibited* in further processing. When a core stereotype underspecifies a linguistic sense given the context, the stereotype will be *enriched*. Importantly, this strategy wouldn't involve changes in the stereotype contents linked to a given concept or linguistic item *per se*. Instead, certain decisions of how to design the linguistic environment in which the item is embedded would merely lead to the retrieved stereotypes to be swiftly enriched or inhibited, and that *modified* stereotype is what influences subsequent on-line linguistic cognition.

While Fischer has certainly developed a promising and interesting workaround for the feasibility problem at issue, this strategy might not satisfy those that are invested in the program outlined at the beginning of this paper: the project of changing problematic ordinary category representations of social groups. This is because Fischer seems to focus on ameliorative aims that are related to, but nevertheless importantly different from, the ameliorative aims that are at the focus of this paper. First, while Fischer's approach focuses on linguistic comprehension, our question was whether we can change the default bodies information involved in all sorts of everyday cognitive tasks in which we employ the relevant concepts. These tasks do not only include linguistic comprehension and production, but also social categorization, automatic inductive inference, social reasoning—among others. But Fischer's approach, which appeals to amendments in *linguistic* environments, would not (at least not straightforwardly) extend to these other cognitive contexts.

Second, as pointed out earlier, Fischer provides an ingenious account of how to *circumvent* the influence of stereotypes on further linguistic processing without changing the stereotypes themselves. But some ameliorators might want to do exactly that: change the *very concept* (i.e., the very default bodies of information) that are associated with a social category, and not merely their deployment in certain contexts. Even if we only consider contexts of language comprehension, in order to prevent that problematic stereotypes are *typically and reliably* guiding the interpretation of a social lexical item, it is crucial to change the *very* stereotype associated with the social category. Take, again, the concept PITBULL, and the stereotypical feature +AGGRESSIVE. Many linguistic contexts in which the term "pitbull" is embedded won't be incompatible with the feature +AGGRESSIVE, so this

stereotype component won't be inhibited, and influence further linguistic cognition. The aim of the kind of ameliorative proposals that my paper focuses on is the revision of the very social concept of relevance, such that the relevant stereotype isn't reliably activated by default by the corresponding lexical item.

In fact, even in linguistic contexts that *are* incompatible with the stereotype in question, Fischer's strategy likely wouldn't lead to an inhibition of the relevant problematic stereotypes.[11] Fischer points out that cases that depend on the inhibition of relevant stereotypes can be subject to a 'control gap': when the stereotype associated with a lexical item is very salient, it might continue to influence linguistic comprehension even if it is strictly incompatible with the linguistic context. Plausibly, many problematic stereotypes associated with social groups will be exactly of that type.[12] Thus, not only is Fischer's aim importantly different from the one we focus on in this paper, even if we were to attempt to inhibit cognitive influence of problematic stereotypes without changing the stereotype itself in the way Fischer envisions, by his own lights, this likely wouldn't work for many of the cases we're interested in. Of course, all this doesn't take away from the many ways in which Fischer's approach *can* lead to desired changes in real-life on-line linguistic comprehension, especially when paired with other kind of concepts.

In sum, then, Fischer (2020)'s account doesn't provide the necessary resources to conceptually engineer the very problematic social stereotypes that are based on statistical co-occurrences. This leaves us with our previous upshot: engineering social concepts without corresponding world change is not feasible. Some conceptual-engineering skeptics might conclude from this that conceptual engineering of social concepts is a pointless enterprise altogether. Against this, in the rest of the paper, I argue that this conclusion is unwarranted, and make the case that we have reasons to be optimistic about the prospects of conceptual engineering. This is because concepts don't only encode statistical information about category features, but also 'intuitive theories', or information about how these features are causally related to each other. My main claim going forward is that conceptual engineers should focus on these causal structures as a locus for conceptual activism and change.

11. Note that this doesn't mean his strategy doesn't work in many other cases of interest to conceptual engineers.

12. See, e.g., Osterhout *et al.* (1997); Palomares (2009, 2008).

4 Concepts and Causal Models

A wide range of evidence about our induction and categorization behavior suggests that conceptual features are not only associated with weights, but are also represented as causally related. Thus, frameworks that merely focus on typical features are incomplete.

Various results from key psychological paradigms support the view that concepts encode causal structure. As an example, let us look at Frank Keil's (1992) famous transformation paradigm. In it, perceptual properties of an object of a particular category were modified as to look and behave like a member of a different category. For example, children and adults were told that a horse was made to completely look and behave like a zebra through operation and training by doctors. The task was to judge whether the animal was a horse or a zebra. Children and adults judge that perceptual appearance didn't affect category membership—a horse stays a horse, even if it looks and behaves like a zebra.

The results have been commonly interpreted as supporting *psychological essentialism*. According to psychological essentialism, categorizers attribute some unobservable constancy to the animals that isn't affected by superficial, observable changes (Gelman 2004; Neufeld 2022). Importantly, the hypothesized *causal structure* underlying essentialist concepts directly explains the judgements in Keil's experiments. Superficial features provide diagnostic evidence for an underlying essence that normally causes the features. If the superficial features are a result of external intervention, the inference from surface features to corresponding essence is defeated. We thus continue to rely on the feature we knew was present before the intervention. This 'undoing effect' is a hallmark effect of causal reasoning (cf. Sloman & Lagnado 2005b; Sloman 2005). A considerable amount of other research provides compelling evidence to support the view that concepts not only encode statistical information about category features, but also information about how these features are causally related.[13]

It is essential to take seriously the statistical *and* causal organization of conceptual structure, especially when applied to social categories. For example, pervasive and early emerging stereotypes represent men as having more 'raw brilliance' than

13. See, e.g., Ahn *et al.* (2000); Gelman & Wellman (1991); Gelman (2003); Rips (2001); Carey (2009); Rehder (2003a,b); Rehder & Hastie (2001, 2004); Rehder & Kim (2010); Rehder (2017); Hayes & Rehder (2012); Rehder (2015); Neufeld (2022); Sloman (2005); Lagnado (2021).

women (Leslie *et al.* 2015; Bian *et al.* 2017, 2018; Storage *et al.* 2020; Muradoglu *et al.* 2021). This is highly correlated with gender representation in fields in which 'raw brilliance' is considered to be particularly important (e.g., philosophy). Against this background, suppose we find that your concepts FEMALE PROFESSOR and MALE PROFESSOR encode the features +SMART, +HARDWORKING as typical (i.e. diagnostic and/or common). Does this mean the concepts are identical in this respect? *No*, because the same feature weights are compatible with different causal models. For example, +SMART might be represented as causally dependent on +HARD WORKING in FEMALE PROFESSOR, but not in MALE PROFESSOR. Work by Del Pinal and colleagues (Del Pinal *et al.* 2017; Del Pinal & Spaulding 2018) suggests that exactly this is the case: the bias is causal, and not merely statistical.[14]

Importantly, these causal structures have downward effects in a variety of psychological and/or behavioral domains. Consider our judgements associated with the concepts under discussion in different compositional contexts. If smartness causally depends on hard work in female, but not male professors, *lazy* female professors won't be judged to be smart, while lazy male professors will likely still be judged smart. Since our concept of MALE PROFESSOR doesn't represent smartness as causally dependent on hard work, it is compatible with our concept of MALE PROFESSOR that they can be smart even when they're lazy. How the features in our concepts are causally arranged also has important consequences for our interpretation of evidence and resistance to counterexamples. For example, if your concept models men as highly disposed to be good at math, and you find a group of men that is bad at math, you can keep your original model alive by adding an auxiliary intervention that simply prevents the disposition from being realized. Finally, the causal structures play key roles in explanation. For example, when we generate explanations for women's performance in, say, letters of reference or teaching evaluations, we will tend to refer to hard work rather than brilliance as giving rise to the performance.[15]

14. To be clear, the bias is likely also statistical. The point is that even if it wasn't, the difference in causal organizations still result in bias.

15. See Dutt *et al.* (2016); Schmader *et al.* (2007) for a evidence regarding differing prevalence of "brilliance" and "productivity" in letters of recommendation for men vs. women (although see Bernstein *et al.* (2022) for conflicting evidence). See Del Pinal & Spaulding (2018); Del Pinal *et al.* (2017) for further discussion.

5 Conceptual Engineering: Redux

The insight we derived from the 'Roschean' tradition of the psychology of concepts was that the only way to change concepts is via world-change. Conceived this way, there's no distinctive role for the social-conceptual engineer. In the last section, however, I showed that this tradition overlooks that concepts also encode causal information. In this section, I show that this insight has important consequences for the prospects of conceptual engineering. That is, there are ways to change concepts without having to change perceived world patterns: by changing the representation of causal relations between conceptual features.

In order to illustrate how the insight that concepts encode causal information changes how we should think about the feasibility of conceptual engineering projects, consider, again, the concept PITBULL. The typical feature +DANGEROUS (as well as +MUSCULAR, +BIG HEAD, etc.) can stand in different causal relations within the causal model of PITBULL. For example, one possibility is that the concept PITBULL might represent the feature +AGGRESSIVE as causally dependent on INHERENT PITBULL PROPERTY (see figure 1). Alternatively, we can represent

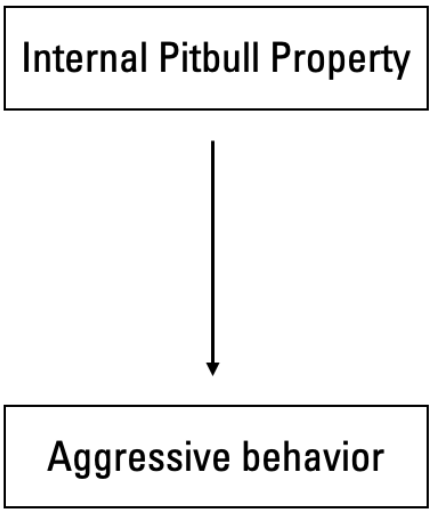

Figure 1: Fragment of a candidate causal model of PITBULL.

INHERENT PITBULL PROPERTY to be causally related to +FRIENDLY. In addition, our

conceptual representation also contains a representation of the feature owned by irresponsible owners, which *causally intervenes* on the variable +friendly behavior, cuts this feature off from its usual causes, and changes its value to +aggressive (see figure 2).

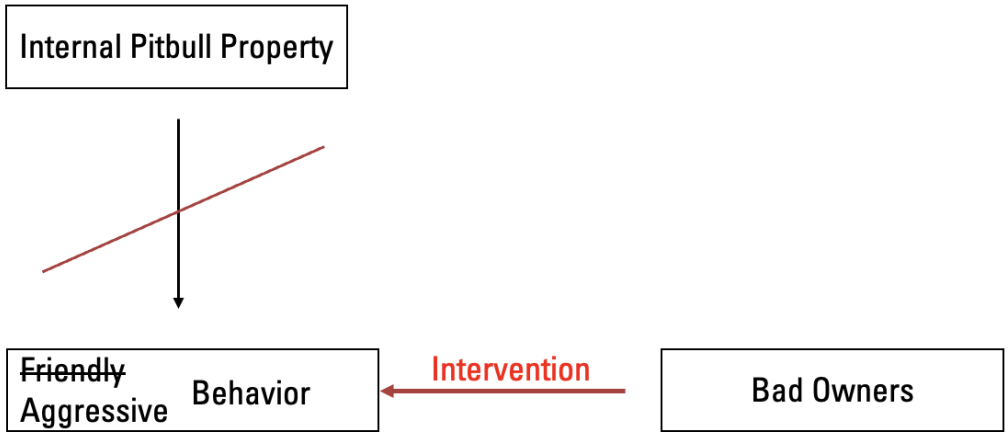

Figure 2: Fragment of a different candidate causal model of pitbull.

Note that for a variety of reasons, the question of which causal model we use to represent pitbulls is quite important. For example, if 'the folk' is predominantly operating with the concept in figure 1, they will more likely push for pitbull breed bans in order to prevent incidents, because this is the policy consistent with their intuitive theory of pitbulls. In contrast, if they operate with the concept in figure 2, the policy changes pushed for will rather focus on dog owners, since they are causally responsible for the undesired behavior. Suppose also that people commonly operate with a concept close to figure 1, but the concept in figure 2 actually corresponds better with reality. In this case, people would push for ineffective policies due to their intuitive theories of pitbulls. This is because a pitbull ban will hardly affect the rate of harmful bite incidents, since bad dog owners will just be able to intervene on the behavior of other dog breeds.

Interestingly, results from cognitive psychology suggest that the intuitive theory encoded in pitbull resembles the one in figure 1. As mentioned earlier, we generally represent natural kinds—including chemicals, plants, and animals—and social kinds—including genders, races, and ethnicities—via a common cause model of the kind posited by psychological essentialism: a hidden, unobservable 'category

essence' serves as intrinsic cause of observable surface features (cf. Gelman 2004; Neufeld 2022). More generally, our intuitive theories seem to be subject to an "inherence bias". This is a "a bias that leads children to overuse intrinsic or inherent features in their explanations" (Sutherland & Cimpian 2019). Suppose you are trying to generate an explanation for the fact that orange juice is a popular breakfast drink. Research suggests that you will prefer an explanation that appeals to internal features of orange juice—e.g., its nutritional content—rather than one that appeals to external features—e.g., an existing orange crop surplus in the 20th century (Cimpian & Salomon 2014; Salomon & Cimpian 2014; Cimpian 2015; Hussak & Cimpian 2015). As a result, not only our concept PITBULL, but also concepts of the social world encode models that skew towards essentialism or give more causal significance to inherent features. In other words, we are disposed to explain, often *inaccurately*, observable surface features that might have structural causes by reference to intrinsic, immutable features (e.g., genes) instead. This, in turn, makes the task of engineering causal components of social concepts particularly important.

The question that's most important for the purposes of this paper is whether conceptual engineering faces the same 'existence threat' it does if the Principles of Categorization exhausted the make up of our conceptual structure. Suppose our target PITBULL concept is the one depicted in figure 1, and we want to revise it as to encode the model in figure 2. As before, we suppose that the typical features +AGGRESSIVE and +MUSCULAR are a reflection of perceived world patterns. Now suppose we succeed and change our concept PITBULL to look more like figure 2. Again, the typical features +AGGRESSIVE and +MUSCULAR adequately reflect the perceived world patterns. Would the conceptual change be threatened by the fact that the perceived world patterns stay unchanged? *No.* This is because the perceived world patterns / data bundles are compatible with the new causal model. In contrast to mere statistical data, causal relations are not as sensitive to mere observational co-occurrences. This is because the same correlational structures are compatible with multiple causal models: a high correlation between lung cancer and yellow teeth can be grounded, for example, in a causal link from yellow teeth to lung cancer, or a common-cause model in which a third variable, smoking, causes both lung cancer and yellow teeth, and is responsible for the statistical correlation. Applied to our example, the same correlational structure between pitbulls and dangerous behavior is compatible with multiple causal models, including the

revised one in Figure 2. Thus—and this is crucial—although relevant world patterns and Principles stay the same, *the new concept can stick*.

6 The Imperative to Engineer

In the previous section, we've seen that there's a way in which conceptual engineers can affect conceptual representations without these changes being undone by input data bundles. Thus, conceptual engineering projects that focus on social concepts have an important function: namely, to find effective ways of changing the *causal structures* we encode in our ordinary concepts of social groups.[16] Interestingly, recent research in developmental and cognitive psychology has demonstrated that even children can flexibly incorporate certain associations between social categories and attributes into different causal models. For example, Ny Vasil and colleagues presented evidence that both children and adults are "able to understand category-property associations (such as the association between "girls" and "liking pink") in structural terms, locating an object of explanation within a larger structure and identifying structural constraints that act on elements of the structure" (Vasilyeva *et al.* 2018, p. 1735).[17] Thus, even if biases towards certain causal structures, such as inherence-based models, exist, research of this kind suggests that a change of how we causally conceptualize social groups by default is in principle possible.

But how do we go about in effecting the relevant conceptual changes in a systematic way? Once the concept is engineered, proposed, and accepted, conceptual engineers are facing the next practical task: they must make concerted efforts and find effective ways to implement the relevant changes. This task will require serious engagement with and utilization of the relevant empirical research on causal and conceptual learning. The key is, however, that the relevant evidence is more sophisticated than mere metrics of probabilistic dependencies. Cognitive scientists have already identified several important cues that both children and adults use to infer the causal structure of a causal system.[18] For example, *temporal information* often serves as a cue for learners. Causes usually come before effects, and learners

16. This also has consequences for some concrete proposals regarding the appropriate targets of conceptual engineering. For example, Isaac (2021a) has proffered that a pluralist model of concepts—both causal *and* statistical—is the appropriate target for conceptual engineering projects. If the line defended in this paper is correct, it means that this proposal should be refined.

17. See also Leshin & Rhodes (2023); Zhang *et al.* (2023) for research on positive cognitive outcomes associated with structure-based conceptual interventions.

18. For an overview, see Lagnado *et al.* (2007).

make use of this (fallible) cue to infer the causal relationship between variables (Lagnado & Sloman 2006). That said, social systems aren't usually the kind of systems we observe unfolding in real-time; thus, this dimension of causal learners might not be the most effective one for affecting the *default* bodies of information we associate with a social category. In addition, higher-order assumptions seem to play an important role for the causal models children and adults form about particular categories. Higher-order assumptions are beliefs such as 'animal's stable behavior has internal causes'. These higher-order beliefs might then influence the default causal models we generate for more specific categories (e.g., pitbulls) (Griffiths *et al.* 2011; Kimura & Gopnik 2019). One way of enacting conceptual change might then be to reform some of the *higher-level* assumptions our category system presumably draws from when constructing causal models about specific social categories.

A particularly important tool for influencing the representation of causal relations, however, is *interventional* evidence (Steyvers *et al.* 2003; Sloman & Lagnado 2005a; Sloman 2005; Lagnado *et al.* 2007; Griffiths & Tenenbaum 2009; Pearl 2009; Pearl & Mackenzie 2018; Bramley *et al.* 2015, 2017). Interventional evidence goes over and above mere observational evidence and allows cognizers to make inferences about causal directions between features. One way of providing interventional evidence is by appealing to deconfounding controls.[19] Suppose our two candidate causal models are salient candidates in our hypothesis space. Subjects can then be presented with scenarios in which the kinds of owners are held fixed across dog breeds. If proportions of, say, aggressive behavior are preserved, we have evidence for the 'inherent' causal model. But if the proportions become indistinguishable, we have shown that type of dog owner is a plausible confounder, and thus presented evidence for a causal model in which dog owners have causal influence on dog behavior.

The fact that correlational (i.e., observational) and causal information are (to an extent) independent can serve to illustrate a deeper point about conceptual engineering. As we saw earlier, a consequence of the Principles of Categorization is that we have to change the world in order to change typical features that we use in categorization and induction. Now, suppose that the world, for whatever reasons,

19. See Pearl *et al.* (2016); Shpitser & Pearl (2008) for a detailed overview of how to perform an interventional *do*-operation via the backdoor criterion—i.e., via blocking spurious paths between two variables. See also Pearl & Mackenzie (2018), ch. 4 for an introductory overview.

changes in the desired ways as to affect statistical associations between a group and typical feature positively. Taking again our PITBULL example, suppose we don't receive <PITBULL, +DANGEROUS> data bundles any more, be it through media, the world, testimony, or other sources. Instead, we are constantly confronted with <PITBULL, +FRIENDLY> pairings. More generally, we can suppose the world undergoes substantial changes such that, due to just systems and structures, our perceived world patterns don't pair social groups with attributes in ways that give rise to statistically-biased conceptual content. In such a scenario, would we still need conceptual engineering?

Yes. This is because the causal model we use to represent a category results in a certain kind of *inertia*: Given that a concept *already* encodes a certain model, we can always *fit* the model to the data. For example, we can add auxiliary hypotheses to make the world patterns consistent with the causal model we started out with (cf. Taylor & Ahn 2012; Lagnado 2021; Sloman 2005; Waldmann 1996). This is important in many respects. When our concept encodes a certain causal model, we will be primed to find evidence that fits the model. But this means that our model would be *constantly* confirmed because we don't realize the compatibility of the data with alternative models—even when the data stem from a more 'just' world. For example, a social group might be associated with seemingly positive typical features, such as +FINANCIALLY SUCCESSFUL and +EDUCATED, but exactly these typical features could be embedded in a causal theory that construes these features as 'unjustly acquired', say, as a result of in-group conspiracies.

Of course, this is not only an abstract possibility, but in many respects captures the actual predicament that we find ourselves in. Given causal biases such as the inherence-heuristic, or independent motivation to attribute certain statistical tendencies to internal causes, a significant proportion of our current concepts of historically marginalized groups are likely to involve models that treat negative statistical tendencies as caused by internal causes. As a result, even when certain properties statistically associated with a social group change for the better, the original causal model *will still determine the way we explain, or explain away, the new associations*. Consider our earlier example of anti-brilliance biases against women. Historically, women have been represented as having less innate intelligence than men. This inherited causal model is likely to affect our interpretation of new evidence: e.g., positive trends in the academic success of women can be at least partly attributed to auxiliary factors such as hard work, so as to preserve models

in which men are portrayed as more likely to be innately brilliant. More generally, evidence provided by the increasing number of women in intellectually-demanding domains can be simply re-interpreted to fit the initial model.

This insight, then, leaves us with a stronger claim. Not only is conceptual engineering possible. In order to achieve certain (conceptual) social goals, it is necessary. There is an imperative to socially engineer. Otherwise, the causal profile we associate with social categories would not only be inaccurate, but is at danger of producing detrimental social, interpersonal, and material consequences.

7 Conclusion

In this paper, I argued that predominant frameworks on the psychology of concepts that shape work on stereotyping, bias, and machine learning inevitably lead us to the view that our categorization and induction practices for social categories can only be reliably changed through world change. Thus, direct conceptual engineering of ordinary social concepts are ineffective, if not impossible. Contrary to this insight, however, I showed that this tradition tends to ignore that concepts don't only encode statistical, but also causal information. Paying attention to this feature of concepts shows, I argued, that there is an imperative to conceptually-engineer, even when the perceived world-patterns change for the better.

Acknowledgments

For helpful comments, conversations, and feedback, I am grateful to audiences at UMass Amherst, University of Toronto, Bowling Green State University, CUNY Graduate Center, Vassar College, the 2023 SSPP in Louisville, the 2022 (E)SPP in Milan, the "Social Identity and Cognition" workshop in Joshua Tree, the Conceptual Engineering Online Lecture Series, the COCOA Zoom Seminar, as well as an anonymous reviewer and an anonymous editor for *Philosophy and Phenomenological Research*. Special thanks to Guillermo Del Pinal for extensive discussions and comments on multiple drafts.

Works Cited

Ahn, Woo-kyoung, Gelman, Susan A, Amsterlaw, A, Hohenstein, Jill, & Kalish, Charles W. 2000. Causal status effect in children's categorization. *Cognition*, **76**, 35–43.

Banaji, Mahzarin R, & Greenwald, Anthony G. 2013. *Blindspot: Hidden biases of good people*. Bantam.

Bernstein, Robert H, Macy, Michael W, Williams, Wendy M, Cameron, Christopher J, Williams-Ceci, Sterling Chance, & Ceci, Stephen J. 2022. Assessing gender bias in particle physics and social science recommendations for academic jobs. *Social sciences*, **11**(2), 74.

Bian, Lin, Leslie, Sarah-Jane, & Cimpian, Andrei. 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, **355**(6323), 389–391.

Bian, Lin, Leslie, Sarah-Jane, Murphy, Mary C, & Cimpian, Andrei. 2018. Messages about brilliance undermine women's interest in educational and professional opportunities. *Journal of experimental social psychology*, **76**, 404–420.

Bloom, Paul. 2014. Can prejudice ever be a good thing? *Ted talk*.

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

Bramley, Neil R, Lagnado, David A, & Speekenbrink, Maarten. 2015. Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of experimental psychology: Learning, memory, and cognition*, **41**(3), 708.

Bramley, Neil R, Dayan, Peter, Griffiths, Thomas L, & Lagnado, David A. 2017. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, **124**(3), 301.

Brownstein, Michael, Madva, Alex, & Gawronski, Bertram. 2019. What do implicit measures measure? *Wiley interdisciplinary reviews: Cognitive science*, **10**(5), e1501.

Burgess, Alexis, & Plunkett, David. 2013. Conceptual ethics i. *Philosophy compass*, **8**(12), 1091–1101.

Cappelen, Herman. 2018. *Fixing language: An essay on conceptual engineering*. Oxford University Press.

Cappelen, Herman. 2020. Experimental philosophy without intuitions: an illustration of why it fails. *Philosophical studies*, 1–9.

Carey, Susan. 2009. *The origin of concepts*. Oxford University Press.

Chomsky, Noam. 1980. Rules and representations. *Behavioral and brain sciences*, **3**(1), 1–15.

Chomsky, Noam, *et al.* 2000. *New horizons in the study of language and mind*. Cambridge University Press.

Cimpian, Andrei. 2015. The inherence heuristic: Generating everyday explanations. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, 1–15.

Cimpian, Andrei, & Salomon, Erika. 2014. The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and brain sciences*, **37**(5), 461–480.

Del Pinal, Guillermo. 2016. Prototypes as compositional components of concepts. *Synthese*, **193**(9), 2899–2927.

Del Pinal, Guillermo. 2018. Meaning, modulation, and context: A multidimensional semantics for truth-conditional pragmatics. *Linguistics and philosophy*, **41**(2), 165–207.

Del Pinal, Guillermo, & Spaulding, Shannon. 2018. Conceptual centrality and implicit bias. *Mind & language*, **33**(1), 95–111.

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**.

Deutsch, Max. 2020. Speaker's reference, stipulation, and a dilemma for conceptual engineers. *Philosophical studies*, **177**(12), 3935–3957.

Díaz León, Esa. 2019. Descriptive vs. ameliorative. *Conceptual engineering and conceptual ethics*, 170.

Dovidio, John F, Kawakami, Kerry, & Gaertner, Samuel L. 2000. Reducing contemporary prejudice: Combating explicit and implicit bias at the individual and intergroup level. *Reducing prejudice and discrimination*, 137–163.

Dutt, Kuheli, Pfaff, Danielle L, Bernstein, Ariel F, Dillard, Joseph S, & Block, Caryn J. 2016. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature geoscience*, **9**(11), 805–808.

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

Fischer, Eugen, & Engelhardt, Paul E. 2017. Stereotypical inferences: Philosophical relevance and psycholinguistic toolkit. *Ratio*, **30**(4), 411–442.

Gelman, Susan A. 2003. *The essential child : origins of essentialism in everyday thought*. Oxford University Press.

Gelman, Susan A. 2004. Psychological essentialism in children. *Trends in cognitive sciences*, **8**(9), 404–409.

Gelman, Susan A, & Wellman, Henry M. 1991. Insides and essences: Early understandings of the non-obvious. *Cognition*, **38**(3), 213–244.

Gonzalez, Antonya Marie, Steele, Jennifer R, Chan, Evelyn F, Lim, Sarah Ashley, & Baron, Andrew Scott. 2021. Developmental differences in the malleability of implicit racial bias following exposure to counterstereotypical exemplars. *Developmental psychology*, **57**(1), 102.

Griffiths, Thomas L, & Tenenbaum, Joshua B. 2009. Theory-based causal induction. *Psychological review*, **116**(4), 661.

Griffiths, Thomas L, Sobel, David M, Tenenbaum, Joshua B, & Gopnik, Alison. 2011. Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive science*, **35**(8), 1407–1455.

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

Hayes, Brett K., & Rehder, Bob. 2012. The development of causal categorization. *Cognitive science*.

Holroyd, Jules, Scaife, Robin, & Stafford, Tom. 2017. What is implicit bias? *Philosophy compass*, **12**(10), e12437.

Hussak, Larisa J, & Cimpian, Andrei. 2015. An early-emerging explanatory heuristic promotes support for the status quo. *Journal of personality and social psychology*, **109**(5), 739.

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

Isaac, Manuel Gustavo. 2021a. Broad-spectrum conceptual engineering. *Ratio*, **34**(4), 286–302.

Isaac, Manuel Gustavo. 2021b. Which concept of concept for conceptual engineering? *Erkenntnis: An international journal of scientific philosophy*, 1–25.

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

Jenkins, Katharine. 2016. Amelioration and inclusion: Gender identity and the concept of woman. *Ethics*, **126**(2), 394–421.

Johnson, Gabbrielle M. 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese*, **198**(10), 9941–9961.

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

Jönsson, Martin L. 2017. Interpersonal sameness of meaning for inferential role semantics. *Journal of philosophical logic*, **46**, 269–297.

Jorem, Sigurd. 2021. Conceptual engineering and the implementation problem. *Inquiry*, **64**(1-2), 186–211.

Kamp, Hans, & Partee, Barbara. 1995. Prototype theory and compositionality. *Cognition*, **57**(2), 129–191.

Keil, Frank C. 1992. *Concepts, kinds, and cognitive development*. mit Press.

Kimura, Katherine, & Gopnik, Alison. 2019. Rational higher-order belief revision in young children. *Child development*, **90**(1), 91–97.

Kitsik, Eve. forthcoming. Epistemic paternalism via conceptual engineering. *Journal of the american philosophical association*.

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

Koch, Steffen. 2021b. The externalist challenge to conceptual engineering. *Synthese*, **198**(1), 327–348.

Lagnado, David A. 2021. *Explaining the evidence: How the mind investigates the world*. Cambridge University Press.

Lagnado, David A, & Sloman, Steven A. 2006. Time as a guide to cause. *Journal of experimental psychology: Learning, memory, and cognition*, **32**(3), 451.

Lagnado, David A, Waldmann, Michael R, Hagmayer, York, & Sloman, Steven A. 2007. Beyond covariation. *Causal learning: Psychology, philosophy, and computation*, 154–172.

Lai, Calvin K, Skinner, Allison L, Cooley, Erin, Murrar, Sohad, Brauer, Markus, Devos, Thierry, Calanchini, Jimmy, Xiao, Y Jenny, Pedram, Christina, Marshburn, Christopher K, *et al.* 2016. Reducing implicit racial

preferences: Ii. intervention effectiveness across time. *Journal of experimental psychology: General*, **145**(8), 1001.

Leshin, Rachel A, & Rhodes, Marjorie. 2023. Structural explanations for inequality reduce children's biases and promote rectification only if they implicate the high-status group. *Proceedings of the national academy of sciences*, **120**(35), e2310573120.

Leslie, Sarah-Jane, Cimpian, Andrei, Meyer, Meredith, & Freeland, Edward. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, **347**(6219), 262–265.

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

Machery, Edouard. 2015. By default: Concepts are accessed in a context-independent manner. *Chap. 20, pages 567–588 of:* Margolis, Eric, & Laurence, Stephen (eds), *The conceptual mind: New directions in the study of concepts*. Cambridge, MA: The MIT Press.

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

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

Machery, Edouard. 2022. Responses to herman cappelen and jennifer nado. *Philosophical studies*, 1–14.

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

Muradoglu, Melis, Horne, Zachary, Hammond, Matthew D, Leslie, Sarah-Jane, & Cimpian, Andrei. 2021. Women—particularly underrepresented minority women—and early-career academics feel like impostors in fields that value brilliance. *Journal of educational psychology*.

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

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

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

Neufeld, Eleonore. forthcoming. Engineering social concepts: Lessons from the science of categorization. *In:* Haslanger, Sally, Jones, Karen, Restall, Greg, Schroeter, François, & Schroeter, Laura (eds), *Mind, language, social hierarchy*. Oxford University Press (OUP). forthcoming.

Nosek, Brian A, Hawkins, Carlee Beth, & Frazier, Rebecca S. 2011. Implicit social cognition: From measures to mechanisms. *Trends in cognitive sciences*, **15**(4), 152–159.

Osterhout, Lee, Bersick, Michael, & McLaughlin, Judith. 1997. Brain potentials reflect violations of gender stereotypes. *Memory & cognition*, **25**(3), 273–285.

Palomares, Nicholas A. 2008. Explaining gender-based language use: Effects of gender identity salience on references to emotion and tentative language in intra-and intergroup contexts. *Human communication research*, **34**(2), 263–286.

Palomares, Nicholas A. 2009. Women are sort of more tentative than men, aren't they? how men and women use tentative language differently, similarly, and counterstereotypically as a function of gender salience. *Communication research*, **36**(4), 538–560.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Pearl, Judea, & Mackenzie, Dana. 2018. *The book of why: the new science of cause and effect*. Basic books.

Pearl, Judea, Glymour, Madelyn, & Jewell, Nicholas P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Quilty-Dunn, Jake. 2021. Polysemy and thought: Toward a generative theory of concepts. *Mind & language*, **36**(1), 158–185.

Rehder, Bob. 2003a. Categorization as causal reasoning. *Cognitive science*, **27**(5), 709–748.

Rehder, Bob. 2003b. A causal-model theory of conceptual representation and categorization. *Journal of experimental psychology: Learning, memory, and cognition*, **29**(6), 1141.

Rehder, Bob. 2015. The role of functional form in causal-based categorization. *Journal of experimental psychology: Learning, memory, and cognition*, **41**(3), 670.

Rehder, Bob. 2017. *Concepts as causal models: Categorization*. Oxford: Oxford University Press. Pages 347–376.

Rehder, Bob, & Hastie, Reid. 2001. Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of experimental psychology: General*, **130**(3), 323.

Rehder, Bob, & Hastie, Reid. 2004. Category coherence and category-based property induction. *Cognition*, **91**(2), 113–153.

Rehder, Bob, & Kim, ShinWoo. 2010. Causal status and coherence in causal-based categorization. *Journal of experimental psychology: Learning, memory, and cognition*.

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

Rips, Lance J. 2001. Necessity and natural categories. *Psychological bulletin*, **127**(6), 827.

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

Rothbart, Myron, & John, Oliver P. 1985. Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact. *Journal of social issues*, **41**(3), 81–104.

Salomon, Erika, & Cimpian, Andrei. 2014. The inherence heuristic as a source of essentialist thought. *Personality and social psychology bulletin*, **40**(10), 1297–1315.

Schmader, Toni, Whitehead, Jessica, & Wysocki, Vicki H. 2007. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles*, **57**(7), 509–514.

Shpitser, Ilya, & Pearl, Judea. 2008. Complete identification methods for the causal hierarchy. *Journal of machine learning research*, **9**, 1941–1979.

Sloman, Steven. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Sloman, Steven A., & Lagnado, David A. 2005a. Do we ?do?? *Cognitive science*, **29**(1), 5–39.

Sloman, Steven A., & Lagnado, David A. 2005b. Do we "do"? *Cognitive science*, **29**(1), 5–39.

Steyvers, Mark, Tenenbaum, Joshua B, Wagenmakers, Eric-Jan, & Blum, Ben. 2003. Inferring causal networks from observations and interventions. *Cognitive science*, **27**(3), 453–489.

Storage, Daniel, Charlesworth, Tessa ES, Banaji, Mahzarin R, & Cimpian, Andrei. 2020. Adults and children implicitly associate brilliance with men more than women. *Journal of experimental social psychology*, **90**, 104020.

Sutherland, Shelbie L, & Cimpian, Andrei. 2019. Developmental evidence for a link between the inherence bias in explanation and psychological essentialism. *Journal of experimental child psychology*, **177**, 265–281.

Taylor, Eric G, & Ahn, Woo-kyoung. 2012. Causal imprinting in causal structure learning. *Cognitive psychology*, **65**(3), 381–413.

Vasilyeva, Nadya, Gopnik, Alison, & Lombrozo, Tania. 2018. The development of structural thinking about social categories. *Developmental psychology*, **54**(9), 1735.

Waldmann, Michael R. 1996. Knowledge-based causal induction. *Pages 47–88 of: Psychology of learning and motivation*, vol. 34. Elsevier.

Weber, Renee, & Crocker, Jennifer. 1983. Cognitive processes in the revision of stereotypic beliefs. *Journal of personality and social psychology*, **45**(5), 961.

Zhang, Marianna Y, Liu, Linda, & Markman, Ellen M. 2023. Let's talk structure: the positive outcomes of structural thinking. *In: Proceedings of the annual meeting of the cognitive science society*, vol. 45.