Conceptual centrality and implicit bias *

Guillermo Del Pinal
ged2102@columbia.edu
ZAS Berlin

Shannon Spaulding
shannon.spaulding@okstate.edu
Oklahoma State University

forthcoming in *Mind & Language*

Abstract

How are biases encoded in our representations of social categories? Philosophical and empirical discussions of implicit bias overwhelmingly focus on *salient* or *statistical associations* between target features and representations of social categories. These are the sorts of associations probed by the Implicit Association Test and various priming tasks. In this paper, we argue that these discussions systematically overlook an alternative way in which biases are encoded, i.e., in the *dependency networks* that are part of our representations of social categories. Dependency networks encode information about how the features in a conceptual representation depend on each other, which determines their degree of centrality in a conceptual representation. Importantly, centrally encoded biases systematically disassociate from those encoded in salient-statistical associations. Furthermore, the degree of centrality of a feature determines its cross-contextual stability: in general, the more central a feature is for a concept, the more likely it is to survive into a wide array of cognitive tasks involving that concept. Accordingly, implicit biases that are encoded in the central features of concepts are predicted to be more resilient across different tasks and contexts. As a result, our distinction between centrally encoded and salient-statistical biases has important theoretical and practical implications.

Keywords: bias; implicit bias; concepts; reasoning; conceptual centrality; essentialism

Words: 9,049

* The authors contributed equally to this work. This paper benefited greatly from the extensive and constructive comments of Alex Madva and two anonymous reviewers for *Mind & Language*. This work was supported by the Alexander von Humboldt foundation.

1 Introduction

Our representations of social categories encode stereotypes and implicit biases that can deeply affect our social judgments and behavior. Some well-known pernicious examples of implicit biases include pairs such as <BLACK, +AGGRESSIVE> and <WOMAN, +BAD AT MATH>. How are such biases encoded in our stereotypes or representations of social categories? Relatedly, are all socially significant biases encoded in the same way or can biases be encoded in fundamentally different ways? Despite the increasing interest in implicit bias,[1] we lack satisfactory answers to these questions. This is surprising given certain trends and tensions in the literature. On the one hand, overviews of the empirical data suggest that various measures of implicit bias tap into different sorts of cognitive processes.[2] On the other hand, most empirical and philosophical investigations proceed as if implicit bias is a single, uniform phenomenon.[3] Specifically, as we show in section 2, philosophical and empirical discussions of implicit bias focus on either *salient* or *statistical* associations between target features and representations of social categories. While there is no doubt that these kinds of associations encode implicit biases and affect judgment and behavior, we argue that other forms of bias encoding may affect social cognition in even more dramatic ways.

The main task of this paper is to show that, due to their narrow focus on salient-statistical associations, discussions of implicit bias systematically overlook an alternative way in which biases are encoded, viz., in the *dependency networks* which are part of our representations of social categories. Dependency networks are structures that capture information about how features in a conceptual representation depend on each other, which in turn determines their degree of centrality. In section 3-4, we show that socially significant biases can be encoded in the dependency networks of our representations of social

1 See, for example, Brownstein and Saul's recently released two-volume collection on the metaphysics and ethics of implicit bias (2016a, 2016b). See also Greenwald et al. (2009) and Nosek et al. (2007) for overviews of the empirical literature.

2 For instance, individual scores on various implicit measures only weakly correlate (Nosek et al. 2007), and in some cases measures of implicit bias correlate with measures of explicit bias and predict overt behavior, but in other cases they do not (Greenwald et al. 2009). This suggests that although the results of these individual tests of implicit bias are robust and reliable, the various tests may be tapping into different cognitive structures.

3 There are some notable exceptions to this trend, e.g., Holroyd & Sweetman (2016). Following Amodio & Devine (2006), Holroyd and Sweetman distinguish implicit semantic associations, implicit affective evaluations, and implicit behavioral motivations as fundamentally different kinds of implicit bias. We will offer a different way to taxonomize implicit bias. Our main focus is not on differences in the nature of the relata, but on differences in the kinds of relations.

categories. Crucially, these biases can be disassociated from those encoded in salient-statistical associations. In other words, there are important forms of social bias that may not show up in measures of salient-statistical associations, but which nevertheless exert a significant influence on judgments and actions. As we argue in section 4, salient-statistical and centrally encoded biases are predicted to behave differently in many cognitive tasks. Specifically, the degree of centrality of a bias determines its cross-contextual stability.

The view that significant social biases can be encoded in dependency networks has important empirical and philosophical implications. We discuss these in sections 5-6. On the empirical side, our notion of centrally encoded biases can shed light on some perplexing patterns in current results, such as the cross-contextual instability of implicit biases and the lack of correlation between different measures and experimental manipulations in studies of bias. On the philosophical side, our account suggests new ways to approach foundational questions, such as whether implicit biases are beliefs. Ultimately, our aim is to provide a more fine-grained taxonomy of socially relevant biases that will serve as a useful starting point for future empirical and philosophical investigations.

2 Implicit bias and salient-statistical associations

Most psychologists and philosophers would agree that, broadly speaking, implicit biases rest on associations between features and our representations of social categories. In this section, we argue that the kinds of implicit biases usually investigated by psychologists and discussed by philosophers are *salient-statistical biases*. The characteristic property of this class of biases is that they depend on *salient* or *statistical* associations between features and representations of social categories. As we will see, current measures of implicit bias are variations of standard measures of salient-statistical associative strength.

Explicit intergroup bias consists in conscious, reflectively endorsed evaluations of social groups. For instance, people who endorse the idea that women naturally are less intelligent than men have an explicit bias against women. Although some people have and are comfortable expressing explicit biases, many people sincerely disavow them. However, a wealth of empirical evidence suggests implicit bias is extremely widespread even amongst those who reflectively endorse egalitarian ideas (De Houwer et al. 2009). *Implicit* biases are representations or evaluations of social groups that occur spontaneously,

are difficult to bring under reflective control, and are sometimes opaque to introspection.[4]

Testing for explicit bias is relatively straightforward. You can just ask people what they think about various social groups and try to control for social desirability censorship. Testing for implicit bias is trickier because, for the most part, people either are unwilling to report or are not consciously aware of implicit processes, and thus you cannot simply ask them about their implicit biases. Instead, experimenters construct tasks that are designed to elicit behavior that is sensitive to such processes, and from the elicited behavior they estimate subjects? implicit bias. For instance, experimenters measure how quickly and accurately subjects associate a social category (e.g., WOMAN) with a feature (e.g., +NURTURING) or the extent to which a priming stimulus representing the social category facilitates a response involving the target feature.

What kind of information is encoded in these sorts of associations between features and our concepts or representations of social categories? And how, precisely, are these associations measured? To answer these questions, it is useful to review some key measures and results in the psychology of concepts. Many psychologists hold that concepts, including those relevant to social cognition, are encoded as sets of weighted features, often called prototypes (Rosch 1975, Rosch & Mervis 1975). Prototype theory offers a useful way to begin to understand the ways in which features can be associated with socially relevant categories. On this view, a concept C typically includes features, $f_1 \ldots f_n$, and each feature is assigned a weight, w_i. There are different accounts of what determines the weight of a feature. A standard view is that weights are a function of statistical properties, such as cue validity, and saliency properties, such as availability and prominence (Hampton 2006, Machery 2006, Morewedge & Kahneman 2010, Murphy 2002).[5] The cue validity of a feature f for a category C is the probability that some entity x belongs to C given that x has f. The notion of saliency is less precise, but it is usually taken to capture a signal-to-noise ratio, such as the prominence of a feature f for category C,

4 Implicit biases, as traditionally conceived, result from learned associations between features and social categories, and they influence our cognitive, affective, and behavioral responses. Though explicit and implicit biases should be distinguished, they can of course align (see Holroyd 2016). For example, one may be both explicitly and implicitly sexist insofar as one reflectively endorses sexist ideas and habitually and reflexively thinks, feels, and behaves toward women in sexist ways.

5 We should note, however, that there are richer notions of prototypes in the psychological literature. For example, Hampton (2006) argues that prototypes often also encode additional structural relations between features and dimensions, such as dependency and degree of centrality. We discuss these kinds of relations between features in the next sections.

or the availability of f in response to names or instances of C. To illustrate, the black and white stripes of zebras have a high cue validity and saliency, and as a result they are a highly weighted feature of our conception of zebras. This is not the case for the feature +HAS HAIR. Accordingly, we expect that priming studies should find a significantly stronger effect between <"zebras" and "stripes"> than between <"zebras" and "hair">. The point here is that degree of association can often be understood as a function of cue validity and saliency. We will call this subclass of associations salient-statistical associations. Table 1 below summarizes the main kinds of salient-statistical associations, and briefly presents an example of the corresponding experimental measures.[6]

Table 1.

Relation/term	Types of measures	Examples
Saliency: Signal-to-noise ratio	Prominence	How prominent in your conception of lions is it that they have manes?
	Availability	Lions have manes (yes/no)
Statistical: Evidence provided by a feature for a category	Cue validity: P(category | feature)	Of all things that have manes, what percentage are lions?
	Typicality: P(feature | category)	Of all the lions, what percentage has manes?

Going back to measures of implicit bias, we can now describe in more detail what sorts of associations between features and concepts of socially relevant categories they uncover. The main point we want to make is that we can

6 A clarification is in order. The main reason why we present measures from prototype theory is to taxonomize 'associations' in terms of the kinds of information they encode. The kinds of information are operationalized in terms of measuring paradigms, as summarized in Table 1. We are not committed to any specific account of the nature of the mental structures that underlie these kinds of associations. (Although we do think that, ultimately, our account provides some useful constraints and suggestions for tackling these questions, as we discuss in §6.) Accordingly, when we talk of the 'prototype' of concept C, we simply mean the set of features that are available, cue-valid, typical, etc. for C. When the claims are comparative, we are usually interested in results such as that f is more available for C_1 than for C_2. These comparative results can be crucial even if the association between, say, f and C_1 is not strong in absolute terms. Furthermore, for our purposes, we need not place any restrictions on the nature of these features (e.g., whether they are amodal, modal, or mixed representations). Finally, different theories of concepts, not to mention different implementation accounts, can account for these sorts of relations in different ways. For interestingly different accounts, see Lakoff (1987), Fodor (1998), and Prinz (2002).

understand the target associations as encoding properties such cue validity and saliency. The reason is simple: most empirical studies of implicit bias are engineered to uncover salient-statistical associations such as those in Table 1. To see this, consider the main experimental paradigms for investigating implicit bias: the Implicit Association Test (IAT) and priming tasks. The IAT measures how quickly and accurately subjects categorize stereotypic and counter-stereotypic associations (Greenwald et al. 1998, 2009). In one well known version of the IAT, subjects are instructed to categorize as quickly and accurately as possible faces of Black men with pleasant words (e.g., "joy", "love", "peace") and faces of White European American men with unpleasant words (e.g., "agony", "terrible", "horrible"). Subjects are then instructed to categorize the stimuli according to the opposite rule: Black faces with unpleasant words and White faces with pleasant words. If subjects categorize faster and more accurately according to one of these rules, they are said to have an implicit bias. As it turns out, most White Americans more strongly associate Black with unpleasant words and White with pleasant words than Black with pleasant and White with unpleasant. For our purposes, what matters here is that IAT is measuring salient-statistical associations, in particular availability. When subjects more quickly and accurately categorize according to one rule, the association that that rule represents is more available to them. Thus, the data show that, for White American participants, categorizing someone as Black makes the unpleasant features more salient or available than when categorizing someone as White.

A second method for investigating implicit bias is through priming tasks which measure the effects of subtle cues in the environment on our emotional and cognitive responses (DeCoster & Claypool 2004, Fiske & Taylor 2013).[7] Subliminal priming occurs when a stimulus is presented to subjects too quickly to be consciously processed. Conscious priming occurs when the subject consciously perceives the prime but has no awareness of its effects on subsequent reactions. The most relevant kind of priming task for our purposes measures cognitive priming, which occurs when subtle cues in the environment activate concepts and influence subjects' judgments. For example, Graham & Lowery (2004) report that police officers and juvenile probation officers subliminally primed with words related to the racial category Black are more likely to interpret a hypothetical adolescent (whose race is unspecified) as having a worse personality, being more blameworthy, more likely to reoffend, and they recommended harsher punishments. Importantly, these priming tasks measure

7 In addition to affective and cognitive priming, many studies report evidence of behavioral priming, e.g., Bargh et al. (1996). The data for behavioral priming are mixed and complicated, and for these reasons we will not discuss them here.

the strength of subjects' salient-statistical associations between a racial category and various features. Put in our terminology, the Graham and Lowery study finds that the subjects in the experiment have a representation of Black people in which delinquency is a prominent feature.

We have just seen that the kinds of associations measured by current empirical research on implicit bias are salient-statistical associations, based on variations of the measures outlined in Table 1. As a result of the overwhelming focus on these measures within social psychology, philosophical accounts of implicit bias also tend to focus on these kinds of associations. Indeed, some accounts focus on these associations to such an extent that it seems fair to say that they tacitly assume that all significant social biases are encoded in salient-statistical associations between features and our representations of categories. We think this is a fundamental mistake.

Assume for a moment that salient-statistical associations are fully determined by properties such as cue validity and saliency and that we know the full associative profile of concepts C_1 and C_2. In other words, we know what features constitute or are associated with C_1 and C_2, and know, in each case, the full salient-statistical profiles of those features. Does it follow that we have all the information we need to determine all significant biases? We clearly have information to determine some biases. For example, suppose our investigation is about POODLE vs. PIT BULL and that the results of an IAT or priming task indicate that the feature +AGGRESSIVE is more strongly associated with PIT BULL than POODLE. If, in fact, poodles are just as aggressive as pit bulls, we can take this pattern of results as evidence of a bias against pit bulls. However, suppose that measures of salient-statistical associations suggest that +AGGRESSIVE is associated with roughly equal strength to POODLE and to PIT BULL. Should we conclude that there is no bias? No. Independently of standard worries about experimental design and drawing conclusions from null results, there may be substantial biases that these measures simply are not designed to detect.

We shall argue below that social biases can be encoded in our concepts in ways that systematically disassociate from salient-statistical properties.[8]

8 Some philosophers object to the identification of concepts with representations that encode features that do not strictly determine their extensions. On this view, mental representations such as prototypes may be part of our conceptions of categories, but in general they are not strictly part of our concepts. To be sure, most philosophers who defend this idea, including Rey (1983), Burge (1993) and Fodor (1998), do not reject the psychological reality of prototypes: in one form or another, they acknowledge that prototypes are key components of our best empirical models of categorization and induction. In this paper, we will employ the term "concept" in the wider, psychological sense. Philosophers who prefer to use the narrower sense should interpret our discussion as being, for the most part, about conceptions

Furthermore, the basic properties of these biases suggest that they likely have a substantial and pervasive effect on everyday social cognition. If we confine our investigations of bias to those that use salient-statistical measures, we will continue to overlook this important class of biases.

3 Conceptual centrality and implicit bias

Concepts, we have suggested, can be represented as sets of weighted features that encode information like cue-validity and saliency. We argued that the kinds of implicit biases currently investigated can be understood in terms of those structures. However, most scientists and philosophers who work on concepts argue that they also encode other kinds of information, in particular, the degree of centrality of their associated features.[9] Our task here is not to directly defend the idea that concepts encode centrality. The relevant results are widely accepted, even by proponents of more recent versions of prototype theory (Hampton 2006). Instead, we aim to (i) spell out the ways in which important social biases can be encoded in the structures that determine centrality, (ii) show that, in general, these biases should be distinguished from those encoded in salient-statistical properties, and (iii) discuss some of the unique properties of centrally encoded biases.

There are various formal models of centrality. We will adopt an account that is intuitive, easy to represent, and has a proven empirical track record. To say that a feature f is central in concept C is to say that other features depend on f more than f depends on them (Sloman & Lagnado 2015, Sloman et al. 1998, Thagard 1989). The notion of dependence is abstract and ranges over more specific dependencies such as causal and explanatory relations.[10] To illustrate, take the concept CHAIR and its features +HAS A BACK and +USED

rather than concepts. Using this terminology, the theme of this paper is about how our conceptions of social categories encode biases, and how that in turn affects cognitive processes such as categorization and induction.

9 Many philosophers argue that concepts encode something like centrality, partly because this provides concepts with stability across informational contexts and cognitive tasks (Putnam 1992). Since the 1980s, psychologists have developed essentialist theories according to which conceptual structures encode the degree of centrality of features (Gelman & Wellman 1991, Keil 1989).

10 There are two ways of interpreting the claim that dependency relations are abstract. (i) We can hold that certain levels of processing (e.g., fast, intuitive processes) are sensitive only to relatively abstract and content-less asymmetric dependencies. (ii) We can say that most levels of processing are sensitive to more specific dependency relations, say, to causal dependency, but that our theoretical claims do not depend on those details. For our purposes, we can leave this issue unresolved, but a full account of the behavior of centrally encoded biases should ultimately address this question.

FOR SITTING. Intuitively, what explains why chairs have a back is that they are designed for sitting. Other typical properties of chairs, e.g., concerning their height and materials, are explained by the fact that chairs are normally designed to be used for sitting. At the same time, these properties are relatively independent of the fact that chairs also normally have a back. It follows that more features depend on +USED FOR SITTING than on +HAS A BACK, and hence the former is a comparatively more central feature of CHAIR. Note that centrality is a matter of degrees. In particular, a feature can be central without being an essence in the traditional sense, and the high centrality of f in C does not entail that it is necessary that every C is f.[11]

To understand the way in which biases can be encoded in dependency networks, we first need to explain the distinction between the degree of centrality of features and their salient-statistical associative properties. The crucial point to make is that centrality and salient-statistical properties disassociate (Sloman et al. 1998).

- Feature f can be central in C and not have either high cue validity or high saliency for C. For example, +HAS A HEART is a central feature of tigers. However, it does not have high-cue validity because so many non-tigers also have a heart. This feature also is not salient because, in the usual encounters, we cannot use it to pick out tigers.

- Feature f can be have high-cue validity or saliency for C and not be central. +YELLOW is a typical and distinctive feature of taxicabs in the United States. However, it is not central because other important features of cabs do not depend on their being yellow as opposed to some other color that could stand out in the relevant ways.

These intuitive examples illustrate dissociations that have been systematically and empirically established. Consider the main results of Sloman et al.'s foundational paper on conceptual centrality. Assume that features $f_1 \ldots f_n$ are the constituents of concept C. Sloman and colleagues show that different measures of centrality correlate in their ordering of $f_1 \ldots f_n$, but do not correlate with any of the orderings determined by measures of either cue validity or saliency. In general, even if we show that feature f is strongly associated with C using experimental paradigms that are sensitive to salient-statistical properties

11 This basic framework is compatible with various versions of essentialism. One could, for example, stipulate that for a feature f to be essential for C just is for f to have a high degree of centrality for C. Our claim that biases can be encoded in dependency relations does not commit us to even this version of essentialism. For as we will show, biases can be encoded in dependency networks even when they do not have a high degree of centrality.

(e.g., IAT and priming paradigms), it does not follow that f is particularly central in C. Similarly, even if we show that f is central in C according to measures that are sensitive to dependency structures, it does not follow that f has a high salient-statistical association with C.

In investigations of biases relevant to social cognition, theorists often are interested in establishing whether feature f plays a different role in concept C_1 vs. C_2. The previous observations entail that even if measures of associative strength indicate that f has the same role (e.g., the same saliency or cue validity) for concept C_1 and C_2, it does not follow that f plays the same role in a broader sense. In particular, f may still substantially differ with respect to its role in the dependency networks of C_1 vs. C_2, and in its corresponding degree of centrality in each case.

To illustrate, suppose we are trying to determine the role of +HAS A BACK in the concepts OFFICE CHAIR and BREAKFAST CHAIR. Suppose we get measures for the saliency and cue validity of +HAS A BACK in each case, and the scores are not significantly different. Given the lack of correlation between salient-associative properties and the degree of centrality of features, this pattern of results would leave wide open the possibility that +HAS A BACK still is more central for one of these concepts. For example, suppose that OFFICE CHAIR has the feature +USED TO SIT FOR MANY HOURS, which itself is relatively central and dependent on +HAS A BACK. Assuming there are no relevant additional differences in the dependency network for +HAS A BACK in each concept, it follows that this feature would be more central for OFFICE CHAIR than for BREAKFAST CHAIR. This holds despite the fact that salient-associative measures would not detect that difference.

When we consider concepts relevant to the study of biases—e.g., concepts of gender, race, ethnicity, and different social roles—the distinction between a feature's salient-statistical scores and its degree of centrality becomes crucially important. For example, an influential recent study shows that women are less represented in academic professions whose members think that success in the field is more dependent on raw brilliance than on hard work or discipline (Leslie et al. 2015). The authors argue that one factor that likely is responsible for that distribution is the existence of gender stereotypes that encode the belief that women posses less raw brilliance than men. Suppose we want to directly investigate whether there is in fact such a 'gender-brilliance' bias in a given social group. It should now be clear that a basic question we have to address is this: How would this stereotype be encoded in the representations of female vs. male academics or potential academics? In light of what we have said so far, it should now be clear that there are two possibilities to consider (not necessarily mutually exclusive): (i) the gender bias is encoded in

salient-statistical associations, or (ii) it is encoded in networks of dependency relations.

Focusing on (i), the hypothesis that most people implicitly or explicitly believe that males are more likely than females to be naturally brilliant may be revealed using measures of saliency-statistical associations, e.g., in priming tasks or more direct tasks that ask participants to estimate the proportion of female professors who are brilliant and the proportion of male professors who are brilliant. Suppose, however, that various measures of the perceived distribution and saliency of brilliance in the relevant representations of male and female categories are indistinguishable. In this case, some theorists might be tempted to conclude that we simply have mistaken pre-theoretic intuitions about our society's stereotypes.

However, the considerations we have raised here suggest that, even if we get those results, we still have to examine option (ii), namely, whether the hypothesized gender-brilliance stereotype is encoded in the relevant dependency networks. Here is one way in which this could happen. Suppose that participants implicitly or explicitly believe that most professors, male or female, are smart and hardworking. This is perfectly compatible with the feature +HARD WORKING being more central for FEMALE PROFESSOR than for MALE PROFESSOR, a result that can obtain if the dependency of +SMART on +HARD WORK is stronger for FEMALE PROFESSOR than for MALE PROFESSOR.[12]

To generalize, a concept C_1 encodes a relative bias if the bias is encoded either in the salient-statistical properties of the constituent features or in the dependency networks that connect those features. Most empirical and theoretical studies of biases in social cognition ignore the second possibility. We have focused on cases in which measures of salient-statistical associations assign feature f a similar score for C_1 as for C_2, even though f is more central for C_1 than for C_2. These cases are useful to emphasize the point that our empirical search for biases relevant to social cognition should not stop at measures of salient-statistical associations, and our philosophical discussions should not be limited to just those kinds of associations. However, we should note that determining the degree of centrality of features is important even in cases where associative scores do differentiate between the role of f in C_1 and C_2. For as we argue in section 5 below, the degree to which a bias is central, and indeed the particular properties of the network which encodes it, allows us to make substantially different behavioral predictions from those we can make if we know only the bias' salient-statistical associative strength.

12 For an empirical exploration of this conjecture, see Del Pinal, Madva & Reuter 2017.

4 Measuring dependency networks and conceptual centrality

We have argued that social biases can be encoded in the dependency networks that determine the degree of centrality of conceptual features. We shall now describe some experimental measures of centrality and briefly discuss how they can be adapted to the study biases. This is an important task. One reason that most empirical studies of implicit bias focus on salient-statistical associations is that measures such as IATs are widely accessible, have provided us with massive amounts of data, and arguably overcome the problem self censorship faced by studies of explicit bias. So, even if we grant, on theoretical grounds, the existence, uniqueness, and potential impact on cognition of centrally encoded biases, we must also show that there are useful measures to discover these biases.

Fortunately, there are a variety of reliable measures of feature dependency and centrality. Sloman, Love & Ahn (1998) propose and empirically substantiate the candidates in Table 2 below. To understand the rationale for these measures, recall that a feature's degree of centrality for a concept C is a function of which other constituent features of C depend on it. It follows that the more central a feature is in C, the greater the impact that eliminating that feature has on the rest of C. To illustrate, suppose that for CHAIR, +SEAT is more central than +FOUR LEGS. Given that more features depend on +SEAT than on +FOUR LEGS, eliminating the former would affect more features of CHAIR than eliminating the latter. Following this basic observation, Sloman et al. propose various paradigms that determine the centrality of a feature f for a concept C by measuring the mutability of f, i.e., the impact that eliminating f has on C. One simple measure is called "Ease-of-imagining". Participants are asked to perform tasks like imagining a real chair that does not have a seat, and to rate the difficulty of doing so. For most participants, it is significantly harder to imagine a chair without a seat than one without four legs. This is not surprising: their use for sitting is an important feature of chairs, and there are many ways to keep that function without having four legs, but it is harder to think of an artifact that can be used for sitting but does not have a seat.

Table 2.

Relation/terms	Types of measures	Examples
Conceptual centrality	Surprise	How surprised would you be to encounter a real chair that does not have a seat?
	Ease-of-imagining	How easily can you imagine a real chair that does not have a seat?
	Goodness-of-example	How good an example of chair would you consider a chair that does not have a seat?
	Similarity-to-an-ideal	How similar is a chair that does not have a seat to an ideal chair?

Importantly, Sloman et al. show that the way in which the four measures of centrality in Table 2 order conceptual features (i) strongly correlate with each other, and (ii) do *not* correlate with the order determined by standard measures of either statistical properties (including cue validity) or saliency (including availability and prominence). Furthermore, Sloman et al. test the claim that degree of centrality can be reduced to position in an asymmetric dependency network. For each target concept, participants are given a picture that contains randomized arrays of its main features, and are asked to draw asymmetric dependency lines to connect them in ways that reflected their intuitions of feature dependencies. For example, given a picture that includes the features +FOR SITTING and +HAS A SEAT, participants could draw an arrow from the former to the latter. Using a simple iterative linear equation, Sloman et al. could then compute the average centrality of each feature for a given concept. The striking result is that these dependency graphs order conceptual features in ways that strongly correlate with the measures of centrality summarized in Table 2, but that, again, do not correlate with the order determined by any of the salient-statistical measures, including those in Table 1.

We have, then, reliable and consistent measures of conceptual centrality. Furthermore, we believe that, with a bit of creativity, these measures can be adapted to investigate centrally encoded biases. Consider how the Ease-of-imagining paradigm could be used to investigate the brilliance-gender bias postulated by Leslie et al. (2015). Suppose we want to test the hypothesis that this bias against the 'natural' brilliance of women is encoded in dependency networks. Using this paradigm, we can ask participants to rate the difficulty of imagining, say, a brilliant Harvard professor who not at all hardworking. One condition would involve a female and the other a male professor. Using a between-subject design in which participants just see the female or the male version of the vignettes would reduce the possibility of self-censorship by

using, say, the male condition as an anchor. If we obtain the result that the imagination task is rated as harder in the female than in the male condition, this would suggest that the feature hardworking is more central in the female representation of otherwise indistinguishable brilliant individuals. We can also combine that basic design with more sophisticated measuring techniques. For example, we can use a mouse-tracking design to detect whether subjects are self-censoring their responses. To implement this we could use a forced-choice design in which subjects are asked whether imagining the target scenario is "easy" or "hard". In set up, the cursor is set at a controlled point in the screen, and subjects have to move it to click in the position the response options. The trick is that the overall mouse routes in each condition can be used to determine if subjects correct themselves along the way to their final categorical responses in ways that betray self-censorship or, more generally, the effect of a bias (Freeman et al. 2011).

Additional paradigms to study centrality and dependency can be found in the essentialist literature, especially within the field of developmental psychology (Carey 2009, Johnson & Keil 2000, Keil 1989). Some of these paradigms can be easily adapted to study centrally encoded biases. For example, Del Pinal, Madva & Reuter (2017) adapted a simple version of a causal reasoning task used by Johnson & Keil (2000) to generate central features. To investigate the hypothesis that the brilliance-gender bias is encoded in dependency structures, the authors used, in one study, partial reasoning schemes such as the following:

> Becoming a Professor is difficult.
> Mary/Jack recently became a Professor.
> Therefore, Mary/Jack must be _____.

Unlike standard ways of generating features for concepts, this kind of reasoning scheme cues participants to generate conceptual features that have central explanatory importance. The variation presented here aims to determine (against appropriate controls) whether there is a difference in the sorts of features freely generated by participants to explain success in intellectual professions for female vs. male targets. For example, if +HARDWORK is generated more frequently for the female than for the male version of the scheme, this would suggest that this feature is more central for FEMALE than for MALE PROFESSOR. As in the previous example, this basic paradigm can also be combined with techniques such as mouse-tracking that are more sensitive to processes self-correction. Although we think that simple designs such as those just described are promising, our goal here is not to defend any particular experimental technique or way of controlling for things like self-censorship.

The aim of this discussion of measures of centrality and feature dependencies is just to provide a reasonable amount of support for the view that many of our current experimental paradigms can be used to begin the empirical search for centrally encoded biases.

Now, even if we grant the theoretical possibility that there are centrally encoded biases, and that we can discover them with several well-understood experimental measures, one might still question their overall importance for everyday social cognition for the following reason. Maybe fast, intuitive processes—which presumably make up a substantial part of social cognition—are insensitive to centrally encoded biases because they are insensitive to the dependency networks that encode them. Many scientists argue that we should think of the computational architecture of higher cognition as divided into fast, intuitive, System 1 processes, and slow, deliberative, System 2 judgments (Morewedge & Kahneman 2010, Kahneman 2011, Sloman 2014). According to some interpretations, System 1 processes are associative in the sense that they are sensitive only to salient-statistical structure, and they operate as a function of activation along such associative pathways. On this view, System 1 processes could be thought to be insensitive to dependency networks, hence to feature centrality.

However, the view that System 1 processes are insensitive to dependency structures is no longer empirically defensible. In particular, recent studies show that fast, intuitive judgments are sensitive to causal structures (Sloman 2014, Sloman & Lagnado 2015), an important class of asymmetric dependency relations. Consider one representative example. Having yellow teeth is correlated with lung cancer. Given that fact, consider whether or not you ought to accept the following recommendation: you should whiten your teeth to lower the probability of getting lung cancer. Obviously the answer is no; and, crucially, most people immediately and intuitively conclude that (Hagmayer & Sloman 2009). That answer rests on sophisticated causal information and analysis. If x has yellow teeth and y has white teeth, then x is more likely than y to get cancer. This is because x is more likely to be a smoker, which is the causally relevant variable for cancer. The act of whitening your teeth amounts to an intervention that makes the color of your teeth independent of the causally relevant variable. There is plenty of evidence of this kind to support the view that fast, intuitive decisions are sensitive to at least one type of dependency structures, namely, causal structures. Thus, there is no reason to think that the intuitive processes involved in everyday social cognition are insensitive to centrally encoded biases.

5 Salient-statistical vs. central features across contexts

We have argued that biases can be encoded in the dependency networks of our representations of social categories and presented some promising paradigms for experimentally measuring these biases. We also argued that if we confine our investigations to measures of salient-statistical associative strength, we will continue to systematically overlook centrally encoded biases.[13] In this section, we discuss key differences in the cross-contextual behavior of salient-statistical vs. centrally encoded biases. We shall argue that features connected to a concept C solely via salient-statistical associations are defeasible in response to certain changes in background information and task demands. In contrast, the more central a feature is in C, the more stable it will be in response to similar variations in information and tasks. In general, the degree of centrality of a feature determines its stability across variations in tasks such as social categorization and induction. This difference between merely associative and central features has important consequences for the role of each kind of implicit bias in social cognition, and helps explain otherwise puzzling empirical findings about the weak correlation among different measures of implicit bias.

To begin our argument, we turn to some key results in studies of conceptual combination. Philosophers have argued, and empirical studies have corroborated, that the salient-statistical features associated with concepts are easily dropped when those concepts enter certain combinatorial environments (Barsalou 1987, Fodor 1998, Fodor & Lepore 2002, Hampton 2006, Rey 1983). Suppose that +MANE is a feature of the prototype of LION. It is reasonable to assume that this feature has high cue validity (given a mane, the likelihood that there is a lion is high) and saliency (it is easy to visually pick out manes). Still, this does not entail that +MANE is preserved under even trivial conceptual combinations involving LION. Consider BABY LION, FEMALE LION, and with a dose of imagination, TRIMMED LION. Note that these combinations are straightforwardly intersective: we are moving from the basic level (e.g., LION) to more specific subcategories (e.g., BABY LION). Although conceptual combination is usually studied with linguistic stimuli, similar sub-categorizations below the basic level are obviously common in non-linguistic cognition. Suppose you are interacting with baby lions at a nursery. To guide your thoughts and actions in that setting, you likely would use a subcategory of LION that

13 We are not denying, of course, that biases encoded in the salient-statistical associations play an important role in social cognition. For instance, implicit biases detected by salient-statistical measures can predict variability in how long we speak to someone, how we evaluate job candidates' resumes, our mood when subliminally exposed to faces of different races, and how far away from someone we are likely to sit (Fazio & Olson 54, Greenwald et al. 2009, Lane et al. 2007).

corresponds to something like BABY LION. As in the linguistic case, features that we strongly associate with LION, such as +MANE, or more importantly in this setting, +DANGEROUS, are not inherited into the relevant subcategory of BABY LIONS.[14]

The compositional behavior of merely salient-statistical features sheds light on debates about why scores across various measures of implicit bias only weakly correlate (Nosek et al. 2007).[15] In some cases, measures of implicit bias correlate with measures of explicit bias and predict overt behavior, but in other cases they do not (Greenwald et al. 2009). Sometimes these patterns are invoked to question the validity of such measures. To evaluate those criticisms, however, we must take account of the following consideration. Most measures of implicit bias, we argue, detect salient-statistical associations. If this is correct, then their lack of stability across contexts may be a result of their behaving in the way non-central prototype features generally behave. Specifically, as illustrated in the case of conceptual combination, salient-statistical features often do not survive sub-categorization, and this must be at least partly responsible for the lack of correlation among measures of implicit bias.[16]

The literature on conceptual combination gives us theoretical reasons to expect that implicit biases detected by priming measures and IATs will behave in the ways described above. The following studies on implicit bias empirically demonstrate this pattern of behavior. As is well known, White American participants tend to display significant anti-Black implicit bias on race IATs. (Govan & Williams 2004) report that changing the subcategory of the exemplar

14 To be clear, there are open debates about the extent to which salient-statistical features are compositional, or, more generally, survive sub-categorization. Some argue that they are highly context-sensitive (Connolly et al. 2007, Fodor 1998, Fodor & Lepore 2002, Gleitman et al. 2012). Others argue that some components of prototypes are relatively stable (Del Pinal 2015, Hampton 1987, 2006, Prinz 2012). This debate includes discussions of models that specify the conditions in which the salient-statistical features of concepts are dropped. As will become clear below, our central point does not depend on assuming that prototypes are radically context sensitive.

15 The weak correlation between implicit measures is explained partly by low reliability of individual implicit measures. This is not unique to implicit social cognition; it also is true of implicit memory tests. However, the correlations are weak even if we factor in the reliability of each test, so this statistical fact does not account for all of the variability in implicit measures (Nosek et al. 2007)

16 As an example of this, Olson & Fazio (2003) show that the lack of correlation between IAT and the Bona Fide Pipeline (BFP) priming task is due to different categorization demands. The race IAT forces categorizing by race, whereas the traditional BFP (and many other priming measures) do not explicitly require categorizing by race. Thus, in cases where results from IAT and priming measures do not correlate, part of the explanation involves differential categorization.

in an IAT affects subjects' results. Typically race IATs use generic pictures of Black and White men, or generic Black names (e.g. Tyrone) and White names (e.g., Josh). Thus, in typical race IATs, the categories are simply Black man and White man. Govan and Williams report that when the stimuli represent subcategories—famous and liked Black men, e.g., Michael Jordan, and famous and disliked White men, e.g., Adolf Hitler—racial bias effects are completely eliminated.[17] Put in our terminology, the association between the category BLACK and negative features does not survive sub-categorization into more specific members of the class; participants did not associate those members with negative features.

Govan and William's experiment is somewhat extreme in that it induces participants to shift from representations of basic level or even superordinate categories to representations of specific categories such as FAMOUS BLACK ATHLETE, or perhaps even of individual level categories such as MICHAEL JORDAN. For this reason, one might accept that there was an elimination of features of the prototypes used at the basic level, and still downplay the general relevance of these results. However, studies that involve less extreme sub-categorizations find similar patterns of results. For example, Wittenbrink et al. (2001) report that White subjects exhibit less negativity in response to Black faces when the Black faces are presented in the background context of a church interior as opposed to an urban street corner. In other words, some of the features associated with White participant's representation of the social category BLACK did not survive into the subcategory CHURCH-GOING BLACK MAN, although they did survive into the subcategory STREET CORNER BLACK MAN.

Unlike many studies and meta-analyses of implicit bias, the two studies just described are structured to directly detect the behavior of associations across levels of categorization. These studies confirm our prediction that implicit biases break down in conceptual combination and sub-categorization. Thus, we have further evidence that IATs and priming studies, common measures of implicit bias, detect salient-statistical associations, and these associations behave just as one would expect any salient-statistical properties of concepts to behave.

In contrast, centrally encoded features are predicted to have a quite different compositional behavior compared to salient-statistical associations, including those detected by common measures of implicit bias. In particular, the more central a feature is to a concept, the more likely it is to survive sub-

17 Importantly, and somewhat depressingly, the results were not *reversed*. White participants did not display behavior indicating implicit bias against the famous, disliked White people, like Adolf Hitler.

categorizations and conceptual combinations involving that concept (Hampton 1987, 2006, Murphy 2002). For example, the feature +BORN OF LION PARENTS is a central feature of LIONS (cf. Keil 1989). According to dominant models of concept composition, this entails that this feature will be more stable across composition and sub-categorization than the less central but salient feature +MANE. This seems intuitively correct. For example, note that our mane-less YOUNG LION, FEMALE LION, and TRIMMED LION all clearly do inherit the feature +BORN OF LION PARENTS. Similarly, suppose we are back at the lion nursery, now operating with the more specific representation BABY LIONS, we can probably agree that although we will not be looking out for manes or hiding behind desks from the baby lions, we still would assume that the baby lions were born in the usual way.[18]

At this point, it should be clear that the degree of centrality of features that encode biases is crucially important to determine the biases' wider role in social cognition. Assume f is more strongly associated with WOMAN than with MAN. Suppose we now establish that f is, in addition, highly central to our conception of WOMAN, and in particular, more central than to our conception of MAN. Being central, we expect that f will have the kind of cross-contextual stability characteristic of such features. This means that it will be comparatively more resilient through conceptual combinations and sub-categorizations. The more central the feature, the more likely it is to survive into STRONG WOMAN, SWEDISH WOMAN, or CONGRESSWOMAN. As we have seen, measures of saliency and cue validity can disassociate from centrality, thus it is possible that, independently of the associative strength between f and WOMAN compared to that between f and MAN, f may be more central for WOMAN and hence more likely to be inherited into all the subcategories of WOMAN that we operate with in our day to day social cognitions.

To sum up, what may initially seem like a technical distinction—viz., the different behavior of merely salient-statistical and central features in compositional combinations—turns out to illuminate a general and fundamental difference in the behavior of implicit biases across contexts. This difference, we have seen, has substantial consequences for how these features project across various manipulations of context and affects the wider role of each kind of

18 To be clear, we are not assuming that even highly central features of a concept C are analytic for C. We can imagine fantastic scenarios in which real lions are not born to lion parents. For our purposes, what matters is just the comparatively greater stability of central features.

bias in social cognition.[19] Put another way, one way to predict the stability of biases is to determine the degree of centrality of the features that encode them.

6 Conclusion: Metaphysical and practical implications

We have argued that biases relevant to social cognition can be encoded in the dependency networks of our conceptual representations. These biases have unique properties, including their behavior in composition and sub-categorization. We have seen that although these central biases cannot be directly picked out using measures that track salient-statistical associations, we do have many experimental paradigms to study them. To conclude, we will briefly discuss the implications of our view for current philosophical debates on the nature of implicit bias.

Debate abounds about the underlying nature of implicit bias.[20] Are implicit biases beliefs or belief-like attitudes? Some argue that the implicit biases probed by priming studies and IATs reveal unconscious implicit beliefs (Mandelbaum 2016) or belief-like propositional attitudes (De Houwer 2014, Levy 2015, Schwitzgebel 2013). Others argue that implicit biases are unique kinds of associative states. Tamar Gendler, for example, argues that implicit bias does not fall squarely into any of our existing categories of mental phenomena. She argues that IAT and similar measures track *aliefs*, a unique type of associative mental state. On this view, implicit biases are habitual affective and behavioral responses to stimuli (Gendler 2008).[21] The taxonomy of biases laid out in this paper provides a fruitful way to sharpen this debate.

One thing in common amongst these views about the metaphysics of implicit bias is that they tend to assume that implicit bias is a uniform kind (Holroyd & Sweetman 2016). When distinctions are made, the focus usually is on the nature of the *relata* of a class of biases (e.g., cognitive vs. affective). We have argued, however, that even if we focus just on the class of biases that involve relations between concepts and cognitive features, implicit biases can be encoded in different ways, including salient-statistical associations and

19 A caveat is in order. We showed that features that are merely associative are not strongly compositional and are only weakly preserved in sub-categorization. Our examples focus mainly on features that have descriptive content. We should add that associations that have emotional valence arguably are more stable in composition and sub-categorization. This interesting point is orthogonal to our main argument. For further discussion of compositionality and expressive terms/components, see (Potts 2005, 2007).

20 See the edited volume by Brownstein & Saul (2016a) for an extended discussion of the metaphysics of implicit bias.

21 Machery (2016) offers an alternative to both belief and alief interpretations, according to which IAT and other such tasks measure traits rather than attitudes themselves.

dependency networks. It follows that we should not assume that this broad class of implicit biases has a uniform underlying nature. Thus, when investigating whether implicit biases are beliefs, we should examine each case independently. For, not only may the answer be different for different kinds of implicit bias, each bias may correspond to a different form of belief. Specifically, we should keep in mind that even if a particular implicit bias is underwritten by a belief, the logical form of that belief might be quite different from that of beliefs that underwrite other implicit biases.

To illustrate, consider the unjust bias involving the pairing of <MUSLIM, +AGGRESSIVE>. In light of our taxonomy, we should not simply ask whether those who have this bias endorse generic claims such as "Muslims are aggressive". Generics can be interpreted in various ways, and there are more direct and transparent ways to represent the alternative contents of each kind of bias. Candidates include the following beliefs: (i) Muslims typically are aggressive, (ii) compared to other salient alternatives (i.e., other social/ethnic groups), Muslims tend to be proportionally more aggressive (even if most of them are not aggressive), or (iii) aggressiveness explains other typical properties of Muslims or their communities. The logical form of each of these beliefs is crucially different: (i) is about the distribution of aggressiveness amongst Muslims, (ii) is about the cue validity of aggressiveness for identifying a Muslim, and (iii) is about the explanatory role and hence centrality of aggressiveness in explaining other typical properties of Muslims. Clearly, the diagnostics we should use to determine whether someone has each kind of belief should be subtly different as well. For example, a subject with belief (i) is more likely to be surprised with evidence that large numbers of Muslims are pacifist than someone with belief (ii). In other words, statistical evidence of the real distribution of aggressiveness amongst Muslims (e.g., that the overwhelming majority are not aggressive), would more directly challenge someone with a belief of form (i) than of form (ii). Revision of the latter depends on additional information about the rate of aggressiveness amongst Muslims compared to other relevant or comparable social groups.

We have seen that the beliefs that could underlie implicit biases may have substantially different logical forms (which are hidden when we express them with a generic statement). The framework we put forward also allows us to see that beliefs with the logical form of centrality, such as in (iii) above, have uniquely important practical consequences. To illustrate, suppose Mary believes that rollercoasters are dangerous. This will then guide Mary's practical and theoretical deliberations in certain ways. For instance, if Mary also believes that one should not do dangerous things, Mary will tend to object to riding rollercoasters. Now, if the form of this belief is just that rollercoasters are

typically dangerous, then we can easily imagine a scenario in which information of some small improvements to rollercoasters make Mary change her mind and consider them to be safe. However, suppose that Mary believes that rollercoasters are inherently dangerous, in the sense that most of their important features depend on their being somewhat dangerous. Then it is somewhat harder to imagine scenarios that would convince Mary that rollercoasters are not dangerous. The latter kind of belief is more resilient.

The key point here is that determining the degree of centrality of a feature helps us predict to what extent an implicit bias will behave in a way that is analogous to these kinds of resilient beliefs. Recall that the degree of centrality of a feature f for a concept C determines, to a large extent, how stable f is for C across contexts. If f is very central for C, then C will inherit f across many variations in background contexts, including those that induce sub-categorization. This entails that f would behave as if it was encoded by a resilient belief. For example, suppose a measure of implicit associations shows, for group G, that +SMART is more strongly associated with MALE PROFESSOR than with FEMALE PROFESSOR. To determine how stable this differential association is likely to be across contexts, we have to determine whether +SMART is also more central for MALE PROFESSOR. To the extent that it is, +SMART is more likely to be inherited into subcategories of MALE PROFESSOR (e.g., MALE ENGLISH PROFESSOR) than of FEMALE PROFESSOR (e.g., FEMALE ENGLISH PROFESSOR) and is more likely to survive variations of background context (e.g., whether the judgment is made in an English or a Physics classroom setting). To the extent that this holds, group G's bias in favor of MALE PROFESSORS would be underwritten by a resilient and central belief that male professors tend to be smarter than female professors, a finding that would have enormous practical implications.

This brief discussion of the logical form and stability of beliefs and biases suggests that theorists concerned with the metaphysical status of biases should also investigate questions about their degree of centrality. Particular kinds of sensitivity to evidence are important factors in deciding whether a mental state counts as a belief with a certain form. To carry out the relevant investigations, we need different kinds of experiments than those that have dominated the implicit bias literature. Furthermore, those interested in practical issues, including designing efficient interventions, should also care about issues of centrality. It is clear that the predicted resilience of a bias across contexts is of great practical significance. These theoretical and practical implications provide further support for the main point of this paper, namely, that there is a unique class of biases, encoded in the dependency networks of our conceptual representations, that has been systematically overlooked by current empirical

and philosophical work. We hope this discussion will motivate empirical work aimed at uncovering actual instances of dependency-based biases and theoretical work that reflects on both their nature and ethical implications.

References

Amodio, D. M. & P. G. Devine. 2006. Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91(4). 652.

Bargh, J. A., M. Chen & L. Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* 71(2). 230.

Barsalou, Lawrence W. 1987. The instability of graded structure: Implications for the nature of concepts. In U. Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, 101–140. Cambridge, UK: Cambridge University Press.

Brownstein, M. & J. Saul (eds.). 2016a. *Implicit bias and philosophy, volume 1: Metaphysics and epistemology*. New York: Oxford University Press.

Brownstein, M. & J. Saul (eds.). 2016b. *Implicit bias and philosophy, volume 2: Moral responsibility, structural justice, and ethics*. New York: Oxford University Press.

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

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

Connolly, A. C., J. Fodor, L. Gleitman & H. Gleitman. 2007. Why stereotypes don't even make good defaults. *Cognition* 103(1). 1–22. http://dx.doi.org/10.1016/j.cognition.2006.02.005.

De Houwer, J. 2014. A propositional model of implicit associations. *Social and Personality Compass* 8(7). 342–353.

De Houwer, J., S. Teige-Mocigemba, A. Spruyt & A. Moors. 2009. Implicit measures: A normative analysis and review. *Psychological bulletin* 135(3). 347.

DeCoster, J. & H. M. Claypool. 2004. A meta-analysis of priming effects on impression formation supporting a general model of information biases. *Personality and social psychology review* 8(1). 2–27.

Del Pinal, G., A. Madva & K. Reuter. 2017. Stereotypes, conceptual centrality and gender bias: an empirical investigation. *Ratio* forthcoming.

Del Pinal, Guillermo. 2015. Dual content semantics, privative adjectives and dynamic compositionality. *Semantics and Pragmatics* 8(7). 1–53.

Fazio, R. H. & M. A. Olson. 54. Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology* 1. 297–327.

Fiske, S. T. & S. E. Taylor. 2013. *Social cognition: from brains to culture*. Los Angeles: SAGE.

Fodor, Jerry. 1998. *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Fodor, Jerry & Ernest Lepore. 2002. *The compositionality papers*. Oxford: Oxford University Press.

Freeman, J., R. Dale & T. Farmer. 2011. Hand in motion reveals mind in motion. *Frontiers in Psychology* 2. 59.

Gelman, Susan A. & Henry W. Wellman. 1991. Insides and essences: early understanding of the non-obvious. *Cognition* 38. 213–244.

Gendler, T. S. 2008. Alief and belief. *Journal of Philosophy* 105(10). 634–663.

Gleitman, L., A. Connolly & S. L. Armstrong. 2012. Can prototype representations support composition and decomposition. In Markus Werning, Wolfram Hinzen & Edouard Machery (eds.), *Oxford handbook of compositionality*, 418–436. Oxford University Press.

Govan, C. L. & K. D. Williams. 2004. Changing the affective valence of the stimulus items influences the iat by re-defininf the category labels. *Journal of Experimental Social Psychology* 40(3). 357–365.

Graham, S. & B. S. Lowery. 2004. Priming unconscious racial stereotypes about adolescent offenders. *Law and human behavior* 28(5). 483–504.

Greenwald, A. G., D. E. McGhee & J. L Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74(6). 1464.

Greenwald, Anthony G, T Andrew Poehlman, Eric Luis Uhlmann & Mahzarin R Banaji. 2009. Understanding and using the implicit association test: Iii. meta-analysis of predictive validity. *Journal of personality and social psychology* 97(1). 17.

Hagmayer, Y. & S. A. Sloman. 2009. Decision makers conceive of their choices as interventions. *Journal of experimental psychology: General* 138(1). 22–38.

Hampton, James A. 1987. Inheritance of attributes in natural concept conjunctions. *Memory & Cognition* 15(1). 55–71.

Hampton, James A. 2006. Concepts as prototypes. *The Psychology of learning and motivation: Advances in research and theory* 46. 79–113.

Holroyd, J. 2016. What do we want from a model of implicit cognition. *Proceedings of the Aristotelean Society* 2(CXVI).

Holroyd, J. & J. Sweetman. 2016. The heterogeneity of implicit bias. In M. Brownstein & J. Saul (eds.), *Implicit bias and philosophy*, vol. 1, New York: Oxford University Press.

Johnson, Christine & Frank C. Keil. 2000. Explanatory knowledge and conceptual combination. In Frank C. Keil & Robert A. Wilson (eds.), *Explanation and cognition*, The MIT Press.

Kahneman, Daniel. 2011. *Thinking, fast and slow*. London: Penguin Books.

Keil, Frank C. 1989. *Concepts, kinds and cognitive development*. Cambridge, MA: The MIT Press.

Lakoff, G. 1987. *Women, fire, and dangerous things: What categories receal about the mind*. Chicago: Chicago University Press.

Lane, Kristin A, Mahzarin R Banaji, Brian A Nosek & Anthony G Greenwald. 2007. Understanding and using the implicit association test: Iv. In Bern Wittebrink & Norbert Schwarz (eds.), *Implicit measures of attitudes*, chap. 3, 59–102. New York: Guilford.

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

Levy, N. 2015. Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Nous* 49. 800–823.

Machery, E. 2016. De-freuding implicit attitudes. In M. Brownstein & J. Saul (eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology*, vol. 1, New York: Oxford University Press.

Machery, Edouard. 2006. *Doing without concepts*. Cambridge, MA: The MIT Press.

Mandelbaum, E. 2016. Attitude, inference, association: on the propositional structure of implicit biases. *Nous* 50. 629—658.

Morewedge, C. K. & D. Kahneman. 2010. Associative processe in intuitive judgment. *Trends in Cognitive Sciences* 14(10). 435–440.

Murphy, Gregory L. 2002. *The big book of concepts*. Cambridge, MA: The MIT PressMIT Press.

Nosek, B. A., A. G. Greenwald & M. R Banaji. 2007. The implicit association test at age7: A methodological and conceptual review. *Automatic processes in social thinking and behavior* 265–292.

Olson, M. A. & R. H. Fazio. 2003. Relations between implicit measure of prejudice: what are we measuring? *Psychological Science* 14(6). 636–639.

Potts, Christoper. 2005. *The logic of conversational implicatures*. Oxford: Oxford University Press.

Potts, Christoper. 2007. The expressive dimension. *Theoretical linguistics* 33(2). 165–198.

Prinz, Jesse. 2002. *Furnishing the mind*. Cambridge, MA: MIT Press.

Prinz, Jesse. 2012. Regaining composure: A defense of prototype compositionality. In Markus Werning, Wolfram Hinzen & Edouard Machery (eds.),

Oxford handbook of compositionality, chap. 21, 437–453. Oxford: Oxford University Press.

Putnam, Hillary. 1992. Reply to chomsky. *Philosophical Topics* 20. 379–385.

Rey, Georges. 1983. Concepts and stereotypes. *Cognition* 15(1). 237–262.

Rosch, E. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General* 104(3). 192–233.

Rosch, E. & C. B. Mervis. 1975. Family resemblance: studies in the internal structure of categories. *Cognitive psychology* 7(4). 573–605.

Schwitzgebel, Eric. 2013. A dispositional approach to attitudes: Thinking outside of the belief box. In *New essays on belief*, 75–99. Springer.

Sloman, Steven A. 2014. Two systems of reasoning, an update. *Dual-process theories of the social mind* 69–79.

Sloman, Steven A & David Lagnado. 2015. Causality in thought. *Annual Review of Psychology* 66. 223–247.

Sloman, Steven A., Bradley C. Love & Woo-Kyoung Ahn. 1998. Feature centrality and conceptual coherence. *Cognitive Science* 22(2). 189–228.

Thagard, Paul. 1989. Explanatory coherence. *Behavioral and brain sciences* 12(03). 435–467.

Wittenbrink, Bernd, Charles M Judd & Bernadette Park. 2001. Spontaneous prejudice in context: variability in automatically activated attitudes. *Journal of personality and social psychology* 81(5). 815.

Guillermo Del Pinal
Leibniz Zentrum für Allgemeine Sprachwissenschaft (ZAS)
Schützenstr. 18, D-10117, Berlin
ged2102@columbia.edu

Shannon Spaulding
Department of Philosophy
246 Murray Hall
Oklahoma State University, USA
shannon.spaulding@okstate.edu