Stereotypes, Conceptual Centrality and Gender Bias

Guillermo Del Pinal, Alex Madva, and Kevin Reuter

August 1, 2016

Abstract

Discussions in social psychology overlook an important way in which biases can be encoded in conceptual representations. Most accounts of implicit bias focus on 'mere associations' between features and representations of social groups. While some have argued that some implicit biases must have a richer conceptual structure, they have said little about what this richer structure might be. To address this lacuna, we build on research in philosophy and cognitive science demonstrating that concepts represent dependency relations between features. These relations, in turn, determine the centrality of a feature f for a concept C: roughly, the more features of C depend on f, the more central f is for C. In this paper, we argue that the dependency networks that link features can encode significant biases. To support this claim, we present a series of studies that show how a particular brilliance-gender bias is encoded in the dependency networks which are part of the concepts of female and male academics. We also argue that biases which are encoded in dependency networks have unique implications for social cognition.

Keywords: conceptual centrality; implicit bias; prototypes; sterotypes; gender bias
Word count:

1 Introduction

The notion of a stereotype is one of the most important theoretical constructs in social psychology. Key properties of implicit biases relevant to social cognition are often directly explained via properties of the stereotypes which encode them, and the notion is generally thought to be sufficiently robust to partially explain some observed patterns of discrimination (Valian, 1998; Banaji and Greenwald, 2013; Beeghly, 2015; Leslie et al., 2015). For example, in an important recent study of gender inequality in academia, Leslie et al. (2015) show that women are underrepresented in fields whose members believe that brilliance is a more important determinant of success than hard work. How could such 'field-specific beliefs' causally affect gender distribution in academic fields? One way, according to the authors, is that field-specific beliefs interact with an pernicious cultural stereotype according to which women are less innately/naturally brilliant than men. In accounts such as this, what, precisely, is the operative notion of a stereotype? In particular, are there different ways in which stereotypes could encode something like this brilliance-gender bias, each with unique implications for social cognition?

The view that our representations of categories include stereotypes or prototypes—the latter, more neutral term being the preferred term in most technical discussions—has a long tradition in cognitive science (Rosch, 1999, 2011; Fodor, 1998; Murphy, 2002; Prinz, 2002; Pinker, 2007; Hampton, 2006; Machery, 2006). To a first approximation, prototypes are sets of features that we use to represent categories. In most accounts, to say that a feature f is associated with category C, or is part of the prototype for C, is to say that f is typical, cue valid, salient, or available for C. We shall call these sorts of relations, 'salient-statistical' associations. Importantly, theoretical discussions of the structure of concepts, partly due to the influence of psychological essentialism, tend to recognise that this notion of 'prototypes' as bundles of salient-statistical features is, at best, incomplete (Keil, 1989; Gelman and Wellman, 1991; Murphy, 2002; Carey, 2009; Hampton, 2006). At the same time, this simple notion has been adopted in social psychology, especially in discussions of implicit bias. Indeed, the most widely used measure of bias—the IAT (Greenwald et al., 1998)—is a measure of the availability of features for certain categories.

To be clear, it is undeniable that the study of stereotypes as salient-statistical associations, and the measures and techniques which trace them, have played a fundamental role in discovering many important social biases, and shed light on many of the wider implications of implicit bias for social cognition (Fazio and Olson, 54; Greenwald et al., 2009; Lane et al., 2007). Still, we think that an exclusive focus on stereotypes/prototypes as bundles of salient-statistical features will result in a radically incomplete understanding of bias and its role in social cognition. More specifically, we will defend the following two claims. First, there are important biases that depend on other, 'deeper' aspects of concepts and prototypes, some of which can elude detection through associative measures. Secondly, biases encoded in dependency networks, including the degree of centrality of features, have unique implications for social cognition.

Consider again Leslie et al. (2015)'s hypothesis that women are stereotyped as having less innate brilliance than men, which we will call the 'brilliance-gender' bias. If we take prototypes as bundles of salient-statistical associations, their hypothesis is naturally construed as saying that features such as *brilliant* are thought to be more likely or typically found amongst men than women, or are more salient or available when people think of male than when they think of female members of certain groups. It follows that the hypothesised brilliance-gender bias should be revealed by measures of typicality, saliency, and related 'associations'. Suppose, however, that IATs don't find that people associate *brilliant* and synonymous terms more strongly with male than with female categories[1], and that measures of typicality don't find that participants think of women as less typically brilliant than men (cf. Study 1 below). If this

[1]Compare, for example, Di Bella et al. (2016), who found in 2 of 3 studies that, on average, participants did not implicitly associate *male* and *philosophy* on an IAT. They did find, however, that men tended to associate *philosophy* with *male*, while *women* tended to associate it with *female*. Does this suggest that women do not actually stereotype philosophy as *male*, or think that success in philosophy depends on stereotypically male traits? We think not, for reasons we explain briefly.

pattern of results is confirmed, and can at least diminish concerns about design and the interpretation of null results, should we conclude that the brilliance-gender stereotype is not as prevalent and problematic as Leslie et al. (2015) propose?

We should not. To see why, we need to consider the aspects of prototypes ignored by the simplified notion, and draw the implications for alternative ways in which biases such as the brilliance-gender stereotype can be encoded. Most cognitive scientists now hold that prototypes represent, in addition to sets of salient-statistical features, information about certain relations between those features (Sloman et al., 1998; Hampton, 2006). These relations, which we will call 'dependency networks', represent how the constituent features of a concept depend on each other. For example, the concept BIRD includes information that *flying* depends on having *wings*. Dependency networks in turn determine the degree of centrality of features. If more features of prototype C depend on f_1 than on f_2, then f_1 is more central for C than f_2. Importantly, the degree of centrality of features doesn't generally correlate with their salient-statistical associative strength:

- f can be central in C and not have a high salient-statistical rating for C. For example, *has a heart* is a central feature of TIGERS. However, it does not have a high cue validity because so many non-tigers also have a heart, and it is not salient because, in the usual encounters, we cannot perceptually use it to pick out tigers.

- f can have a high salient-statistical rating for C and yet not be central. *Striped* is a salient and typical feature of tigers, useful to pick them out. However, it is unlikely to be central because most features of tigers do not depend on their being striped.

Sloman, Love, and Ahn (1998)'s foundational paper on conceptual centrality provides substantial evidence for this disassociation. Assume that features $f_1 \dots f_n$ are the constituents of C. Sloman et al. show that various measures of centrality correlate in their ordering of $f_1 \dots f_n$, but do not correlate with any of the orderings determined by measures of either typicality, cue validity, saliency, or availability. Hence, even if f is associated with C in terms of salient-statistical associations, it does not follow that f is central for C; and even if f is central for C, it does not follow that f is a salient or typical feature of C.

It should now be easy to see why Leslie et al. (2015)'s brilliance-gender stereotype should *not* be assumed to be primarily encoded in a pattern of salient-statistical associations. Given the lack of correlation just described, even if the brilliance-gender stereotype is not encoded in salient-statistical associations, it could still be encoded in dependency networks. In principle, it is easy to see how this could happen. As target features, take *hardworking* and *smart*, and as target concepts, take FEMALE and MALE PROFESSORS.[2] Suppose *hardworking* and *smart* are judged to be equally distributed amongst female and male professors

[2]The brilliance-gender stereotype could be explored using different features and target concepts. Indeed, it might seem that we should explore stereotypes for, say, FEMALE and

(cf. Study 1 below). This is compatible with *hardworking* being more central for FEMALE PROFESSOR than for MALE PROFESSOR. This could happen if the stereotypes have the following structure: the dependency of *smart* on *hardworking* is stronger in FEMALE PROFESSOR than in MALE PROFESSOR. Intuitively speaking, this would mean that even if female professors are thought to be, on average, as smart as their male colleagues, they are implicitly assumed to have had to work harder for that. This bias could have significant consequences. For example, suppose Paul learns that Professors Peter and Mary are not that hardworking. Because Paul also implicitly thinks that female professors' intelligence depends on their being hardworking, he might conclude that Professor Mary is not that smart. At the same time, because he thinks that male professors' intelligence does not depend on being hardworking (e.g., because they are just innately smart), he draws no similar conclusions about Professor Peter.

Those underlying dependencies, then, represent one way in which the brilliance-gender stereotype could be encoded. Studies 2-3 presented below strongly suggest that the scenario just described is indeed one way in which dependency networks encode the brilliance-gender stereotype. Specifically, the studies, taken together, suggest that (i) dependency networks can encode socially significant biases, (ii) that there are useful measures of centrality and dependency between features, which can be adapted to study biases, and (iii) that we would overlook and misunderstand the nature of these biases if we focused only on measures of salient-statistical associations.

2 Preliminary Study: Semantic Feature Production Task

To determine how the brilliance-gender stereotype is encoded, and what role features such as *smart* and *hardworking* play in that, it is important to first empirically generate the relevant prototypes. As mentioned in the Introduction, the brilliance-gender stereotype could be manifested in various, somewhat different gender concepts, including FEMALE/MALE STUDENT, FEMALE/MALE CHILD, WOMAN, MAN, etc. In these studies, we focus on the concepts of FEMALE and MALE PROFESSORS. Accordingly, the aim of this Preliminary Study was to generate prototypes for those target concepts. Given that information, we can determine whether target features such as *smart* and *hardworking* are in fact part of the prototypes, and whether there are any gender differences with respect to how frequently participants generate those target features. To do this, we used a simple semantic feature production task, which is a standard way of generating prototypes (McRae et al., 2005). This Preliminary Study will give some initial insights into the structure of our target concepts, which we will then explore in detail in Studies 1-3. The prediction is that we should observe some key gender differences, in FEMALE and MALE PROFESSOR, with respect to the

MALE STUDENTS. The reasons why we focused on the class of professor will emerge in the course of this discussion.

generation of the target features. It is important to emphasise that, although this Preliminary Study is not designed to test any fine-grained claims, that is in no way a trivial prediction. For although it is generally assumed that even self-described egalitarian societies suffer from something like a brilliance-gender stereotype, this view could well be mistaken. Indeed, Leslie et al. (2015) assume and do not directly investigate that claim, although of course that is not their main concern.[3] Furthermore, it might be that, even if the brilliance-gender stereotype exists, it is not manifested in the concepts of FEMALE and MALE PROFESSORS.

2.1 Methods

312 participants were recruited through Amazon Mechanical Turk and reimbursed for their participation.[4] Using a between-subject design, participants were asked to generate features for certain social categories. Participants were randomly assigned to a condition ($N = 103$) featuring either a female or a male professor. To determine whether participants were processing the mentioned profession and not just the gender of the target stimuli, and vice versa, we also asked participants to generate features in two control conditions. The first control condition asked participants to generate features for a female or male baker ($N = 101$), and the second for an actress or actor ($N = 108$). The target stimuli read as follows:

> Imagine that Mary/Jack is a professor at a university.
>
> Please list five features that you think are typical of Mary/Jack.

2.2 Results

The key results of the Preliminary Study are summarised in Table 1. The feature which was most frequently produced, for the prototype of both female and male professor, was *smart* and synonymous features such as *intelligent* and *brilliant* (we assume here that these terms form an equivalence class, which we label *smart*). However, the production frequency of *smart* was slightly but not

[3] The aim of Leslie et al. (2015)'s study is to show that field-specific beliefs predict female underrepresentation, and not to directly examine the nature of the hypothesised brilliance-gender stereotype (but see Meyer et al. (2015) and some of the references cited there). At the same time, it is important for their overall story that there actually be something like the brilliance-gender stereotype. First, this stereotype is invoked in their account of the mechanisms that causally connect the field-specific belief with the observed gender distributions across academic fields. Second, the assumption that it exists affects their recommendations for tackling the underrepresentation of women in various fields.

[4] All the participants for our studies were recruited using Amazon's Mechanical Turk. As Meyer et al. (2015) point out, Mechanical Turk offers a convenient sample rather than a fully nationally representative sample: several studies show that women are overrepresented, workers are typically younger and more educated than average, and Blacks and Hispanics are underrepresented (Berinsky et al., 2012; Paolacci and Chandler, 2014). Still, as they point out, 'the diversity of an MTurk sample is arguably higher than that of most samples used in human subjects research (i.e., college samples)...'.

significantly higher in the female (76.6%) than in the male condition (72.4%): $\chi^2 = 1.264, p = 0.473$. This result is unlikely to be due to participants' processing only the gender information, for there are differences in the production frequency of *smart* for the control categories within each gender. Another feature frequently produced for female and male professors is the equivalence class of *hard work* and synonyms such as *dedicated* and *committed*. Importantly, we found that the production frequency of *hard work* was almost twice as high for female (39.6%) than for male professors (21.3%). This difference was significant: $\chi^2 = 4.71, p = 0.030$. As shown in Table 1, there are no significant differences in the production frequency of *hard work* in the female vs. male versions of the control conditions. This suggests that these results are not due to participants focusing just on female vs. male groups, or professors vs. other professions, but specifically on female vs. male professors.

	Male	Female	χ^2	p
Professor +hardworking	21.3%	39.5%	4.709	0.030
Professor +smart	72.4.%	76.6%	1.264	0.473
Actor/Actress +hardworking	7.5%	10.9%	0.36	0.547
Actor/Actress +smart	17%	10.9%	0.77	0.380
Baker +hardworking	24.5.%	19.2%	0.41	0.522
Baker +smart	10.2%	1.9%	3.09	0.079

Table 1: Results of the Preliminary Study in terms of % of participants who generated the feature for the specified category.

2.3 Discussion

This Preliminary Study used a standard semantic feature production task to generate prototypes for FEMALE and MALE PROFESSORS. The results support our choice of stimuli for investigating the brilliance-gender stereotype. First, (the equivalence classes of) *smart* and *hard work* were the most frequently produced features for each class, and hence are clearly constituents of the corresponding prototypes. Second, we observe a key and significant gender difference involving these features, namely, that *hard work* was produced more frequently for FEMALE PROFESSOR. In addition to validating our stimuli, this Preliminary Study also has important implications for certain—and at first glance tempting—views about how that the brilliance-stereotype is encoded. Perhaps the most intuitive prediction would be that the stereotype would be encoded in a 'direct' way, with the production frequency of *smart* being significantly higher for male than for female professors. The results suggest that the brilliance-gender stereotype is encoded in a more intricate way, which is somehow related to the significantly greater weight of *hard work* in FEMALE PROFESSOR. Importantly, this difference is unlikely to be due to a general stereotype according to which female *professionals* are more hard working, since this significant differ-

ence is not observed in the non-academic professions used in our controls.

This Preliminary Study is neither intended nor can be used to determine what the detailed structural role of *hardworking* in the target prototypes might be, and how that relates to the hypothesis that there is a brilliance-gender stereotype. Still, one might be tempted by the following train of thought: female professors are represented as more hardworking than male professors because they are assumed to have to make up for their having less raw or innate brilliance. This view would have to be squared with another key result of this Preliminary Study, namely, that *smart* was produced with very high and equal frequency for female as for male professors. Indeed, it might initially seem that this result is in tension with the hypothesis that there is a brilliance-gender stereotype. This is why it is crucial to remember that, as we argued in the Introduction, prototypes encode not only salient-statistical associations, but also dependencies between features. Even if *smart* is equally typical for FEMALE and MALE PROFESSORS, it might differ in terms of its role in the corresponding dependency networks, and thereby encode a brilliance-gender bias. Studies 2-3 below support and will allow us to elaborate this proposal.

3 Study 1: Typicality Experiments

The Preliminary Study shows that a key difference between the prototypes for FEMALE and MALE PROFESSORS is the significantly higher weight of *hard work* in the former. According to the framework we laid out in the Introduction, to fully understand this difference we have to determine the degree of centrality of *hard work* and its position in the dependency networks for each of the target prototypes. However, it is also important to examine the relation the prototypes generated in the Preliminary Study and direct judgments about the distribution of *smart* and *hard work* amongst female and male professors. We need to do this for at least three reasons. First, it could be that the difference between *hard work* in FEMALE and MALE PROFESSOR is due to differences in participants' estimates of the distribution of that feature in each class. That is, it could be that participants simply think female professors are more likely to be hardworking than male professors. Second, it could also turn out that the semantic feature production task is not sensitive to differences in judgments about the distribution of *smart* amongst female and male professors. In other words, even if there is no difference in production frequency of *smart* for the target gender categories, participants might still judge that, say, female professors are more likely to be smart than male professors, or vice versa. Third, judgments of the distribution of *smart* and *hard work* help us determine whether the differences obtained in the Preliminary Study are due to the perceived statistical properties of these features in the target classes. The aim of Studies 1a-b is to examine these possibilities. As in the Preliminary Study, each participant received questions about only one gender. In this way, participants could not compare their answers across the female/male conditions, and censor themselves by correcting any perceived gender differences.

3.1 Methods of Study 1a

186 participants were assigned to both versions of either the female ($N = 94$) or the male ($N = 92$) questions:

(1) Consider the class of female professors. What percentage of all those professors do you think are very smart/hardworking? Please give your best estimate.

(2) Consider the class of male professors. What percentage of all those professors do you think are very smart/hardworking? Please give your best estimate.

3.2 Results

The results of Study 1a are summarized in Table 2. We analyzed the data using a repeated measures ANOVA with *Gender* a between-subject factor and *Smart/Hardworking* a within-subject factor. There was a significant main effect for *Smart/Hardworking* $F(1, 184) = 66.69; p < 0.001, \eta^2 = 0.27$, but no significant main effect for *Gender*, $F(1, 184) = 0.97; p = 0.324, \eta^2 = 0.01$. A significant interaction was recorded for *Smart/Hardworking*Gender*: $F(1, 184) = 5.23; p = 0.023, \eta^2 = 0.03$. A simple t-test showed that there is a small but non-significant difference in the proportion of female vs. male professors who are believed to be hardworking; $t(184) = 1.804, p = 0.073$. Slightly more female than male professors are believed to be hardworking ($M = 67.5\%, SD = 23.4$ vs. $M = 61.3\%, SD = 23.3$).

	Male Prof.	Female Prof.
hardworking	61.3%	67.8%
smart	76.1%	75.8%

Table 2: Results of Study 1a: listing the response frequencies for male and female professors for *hardworking* and *smart*.

3.3 Methods of Study 1b

Study 1b is a different way of approaching the same issue examined in Study 1a. One worry with Study 1a is that mentioning an abstract category such as 'female professors' might signal to participants that they are engaged in gender task. So in Study 1b we used common first names instead. We presented 104 participants with a female or male professor, and asked them to rate, on a 7-point Likert scale (1= 'not very likely' and 7= 'very likely') one of the following questions (female&smart: $N = 24$, female&hard work: $N = 26$, male&smart: $N = 26$, male&hard work: $N = 26$):

(3) Mary is a professor. How likely do you think it is that she is very smart/hardworking?

(4) Jack is a professor. How likely do you think it is that he is very smart/hardworking?

3.4 Results

The results of Study 1b are presented in Table 3. The results corroborate the results of Study 1a. Participants believed that a female or male professor is likely to be both smart and hardworking, and there was no significant gender difference in the likelihood judgements. A 2 X 2 ANOVA was performed to analyze the data. No significant effects were found for the independent factors *Gender* and *Smart/Hardworking*: $F(1, 103) = 0.05; p = 0.832$ & $F(1, 103) = 1.11; p = 0.295$. The small, non-significant differences obtained in this study are in the same direction as those obtained in Study 2a: men get slightly higher numbers for *smart* and women for *hard work*.

	Male Prof.	Female Prof.
hardworking	5.7	5.9
smart	6.0	5.9

Table 3: Results Study 1b: Mean values for all four conditions.

3.5 Discussion of Studies 1a-b

The results of Studies 1a-b are clear: participants' judgements about the distribution of *smart* and *hard work* amongst female and male professors do not show any significant differences. This lack of effect is interesting in light of the Preliminary Study, which shows that *hard work* has more weight in the prototype for FEMALE than in MALE PROFESSOR. As we argued before, the weight of a feature in a concept is due to various factors, including its degree of typicality and centrality. Studies 1a-b examine whether the differences obtained in the Preliminary Study could simply be due to perceived differences in the distribution of *hard work* in the class of female vs. male professors. Two different, corroborating sources of evidence argue against that possibility. In addition, Studies 1a-b also examine whether, despite the lack of a difference in the free production task, participants still judge that male professors are more likely to be smart compared to female professors. This would be a relatively direct way of encoding the brilliance-gender stereotype, but the results clearly argue against this hypothesis. To be sure, Studies 1a-b use explicit, direct measures, which one might reasonably worry allow for some degree of self-censorship. However, our between-subject design reduces the possibility that the lack of a gender effect is due participants adjusting their estimates to eliminate gender differences. Overall, then, Studies 1a-b suggest that the uniquely high weight of *hard work*

for FEMALE PROFESSORS might be more intimately connected with its degree of centrality and interdependencies than with its purely statistical properties such as typicality and cue validity.[5] Studies 2-3 directly examine this suggestion.

4 Study 2: Centrality via causal reasoning task

We have seen that *hardworking* has more weight in FEMALE PROFESSOR than in MALE PROFESSOR. Studies 1a-b suggest that this difference is not due to participants judging that female professors are more likely to be hardworking than their male counterparts. Now, prototypes, we have argued, encode not only salient-statistical associations, but also information about the interdependency between features, and their degree of centrality. Studies 2 and 3 explore our main hypothesis, namely, that the brilliance-gender stereotype, as manifested in concepts for professors, is encoded in the corresponding dependency networks. We begin by investigating, in Study 2, whether there are differences in the overall centrality of *hard work* and *smart* in FEMALE VS MALE PROFESSORS.

To appreciate the motivation behind Study 2, we must understand why measures of typicality and centrality/dependency can dissociate. Suppose that the prototypes OFFICE CHAIR and BREAKFAST CHAIR both include the feature *has a back*, which is judged to be equally typical. Even so, *has a back* might have a different degree of centrality in each prototype. For example, office chairs are mostly used to sit for extended periods of time. Comfort is very important. Breakfast chairs are mostly used for shorter periods of time. So although comfort is important, other things might also matter, say, being compact. Accordingly, people might think that, even if *has a back* just happens to be found with similar likelihood amongst office and breakfast chairs, it is significantly more central for office chairs. Regardless of the current distribution, compared to breakfast chairs, only the basic function of office chairs directly depends on having a back.

Following this logic, Study 2 examines whether, despite being indistinguishable in terms of their perceived likelihood amongst female and male professors, the features *hardworking* and *smart* differ in their degree of centrality. The design we adopt is based on work by Johnson and Keil (2000).[6] We adapted a simple causal reasoning task in which participants are asked to produce features that they think are causally or explanatorily 'deep'. As in the Preliminary Study, this was a free production task, but in this case participants had to generate features that best 'explain' key properties of the female/male professor target class. Since we are interested in gender differences in the way in which *smart* and *hardworking* are thought to account for academic success, we de-

[5] If the likelihood of f for class C_1 and for class C_2 is the same, then the cue validity of f for C_1 ($= P(C_1|f)$) cannot be different from its cue validity for C_2 ($= P(C_2|f)$), assuming they are compared with reference to the same alternative classes (as is likely the case when comparing FEMALE and MALE PROFESSORS).

[6] Frank Keil, in particular, has developed various experimental designs for tapping into intuitions of the centrality of features for particular concepts (see, e.g., Keil, 1989, 2003, 2006).

signed a scheme that asked participants to generate the features 'best explain' why female and male individuals managed to become successful professors. Our hypothesis predicts that we should observe a significantly higher production frequency of *hardworking* in the causal scheme for female professors.

4.1 Methods

203 participants were randomly assigned to one of four conditions: two target conditions featuring either a female ($N = 51$) or a male professor ($N = 50$), and two control conditions featuring either an actress ($N = 52$) or an actor ($N = 50$). The respective reasoning schemes had the following form:

(5) a. Becoming a professor (actress/actor) is hard.
b. Mary/Jack has recently become a professor (actress/actor).
c. Therefore, Mary/Jack must be _____.

Participants were asked to enter the feature that they think would best fit the reasoning scheme.

4.2 Results

The results of Study 2 are summarised in Table 4. In the female and the male versions of the target condition (*professor*), the features most frequently produced were *hard work* and *smart* (i.e., the equivalence class of synonymous terms in each case). 45.1% of participants produced *hard work* for *female professor* and 27.5% for *male professor*. There was a minor difference in the production of *smart*: 29.4% for *female professor*, and 27.5% for *male professor*. In contrast, in the control condition, 57.7% of the participants produced *hard work* for *actress*, compared to 68.0% for *actor*. No participant produced *smart* for *actress* and only one for *actor*. This pattern of results indicate that participants were processing the stimuli as intended, and took account of the specific profession under consideration: although both professions require hard work for success, being smart is judged to be important for *professors* but irrelevant for successful acting careers.

To examine gender differences, we used a binary logistic regression and compared the target (*professor*) with the control condition (*actor/actress*) as well as the effect of gender. *Hard work* responses were coded as 1s whereas any other responses were coded as 0s. The logistic regression model was statistically significant, $\chi^2(4) = 21.810, p < 0.001$. The model correctly classified 64.0% of cases, Nagelkerke $R^2 = 0.136$. *Gender* (*male, female*), ($B = -0.984, Wald\ \chi^2 = 4.246, p = 0.039$), and *Condition* (*experimental, control*), ($B = -1.880, Wald\ \chi^2 = 17.469, p < 0.001$), were significant predictors. The interaction between *Gender* and *Condition* was also significant ($B = 1.338, Wald\ \chi^2 = 4.985, p = 0.026$). A simple χ^2 analysis revealed that the difference in the production frequency of *hard work* in the experimental condition was significant ($\chi^2 = 4.38, p = 0.036$). We also ran a binary logistic regression for *smart*: *smart* responses were coded as 1s and any other responses as 0s. While the model was

statistically significant ($\chi^2(4) = 39.137, p < 0.001$, Nagelkerke $R^2 = 0.309$) and 95.2% of responses correctly classified, only *Condition* was a significant predictor ($B = -2.976, Wald\ \chi^2 = 7.902, p = 0.005$).

	Male Prof.	Female Prof.	Actor	Actress
hardworking	27.5%	45.5%	68.0%	57.7%
smart	27.5.%	29.4%	2.0%	0.0%

Table 4: Results of Study 2: Production frequency for both the experimental condition (professor) as well as the control condition (actor/actress) for *hardworking* and *smart*.

4.3 Discussion

Study 2 tapped into features that are explanatorily central for FEMALE and MALE PROFESSORS. Specifically, participants generated features to complete a reasoning scheme which 'explained' why female or male individuals managed to become professors. In contrast to the lack of gender differences observed in the typicality Studies 1a-b, participants in Study 2 were more likely to generate *hardworking* for the scheme involving a female professor. This suggests that *hardworking* has more weight in the prototype for FEMALE PROFESSOR because it is more central. The results also support part of our hypothesis, namely, that gender differences are encoded in differences in the degree of centrality of features.[7] Note that our control condition reversed the main result: *hardworking* was generated more frequently for actors than for actresses. Hence, it is unlikely that the main result is due to participants assuming that, in general or for *any* given profession, women have to work harder than men for similar achievements. Overall, the results of the causal reasoning task strongly suggest that *hardworking* has a more central role in the prototypes for FEMALE than for MALE PROFESSOR.

Study 2 has, however, an important limitation. Our hypothesis is not just that there are gender differences in FEMALE vs. MALE PROFESSOR. It is that these differences, as manifested in those concepts, encode something like the brilliance-gender stereotype hypothesised by Leslie et al. (2015), according to which women are represented as less naturally brilliant than men. Now, it might be tempting to connect the results of Study 2 and the brilliance-gender stereotype as follows. Since participants think that women have to work harder than men to reach the same level of academic success, doesn't this reveal an implicit assumption that women have less raw brilliance, which is presumably why they have to work harder? We cannot yet jump to that conclusion. This is

[7] Our stimuli mentioned that becoming a professor is difficult. Due to priming, this might have caused an overall increase in the frequency of production of terms such as *hard work*. However, even if this is case, it does not affect the main result. That priming effect should affect both the female and male conditions, so it cannot account for the significantly *different* frequency with which *hard work* was generated across those conditions.

in part because, in the causal reasoning task, *smart* was produced with high and indistinguishable frequency for FEMALE and MALE PROFESSORS. This can be reasonably taken to suggest that there might be other reasons, not connected with presumed brilliance-gender differences, why participants believe that female professors have to work harder than their male counterparts (e.g., maybe people assume that they simply face more obstacles). In short, despite the observed gender differences in the centrality of *hard work*, we do not yet have direct evidence that the brilliance-gender stereotype is encoded in dependency networks.

5 Study 3: Gender differences in dependency between *smart* and *hard work*

Study 2 shows that *hardworking* is more central in FEMALE than in MALE PROFESSOR. We cannot yet conclude, however, that this is because people implicitly assume an intellectual disadvantage, since it might be due to the assumed presence of obstacles independent of that stereotype. Now, the reason why the role of *hardworking* seems crucial to determine how the brilliance-gender stereotype is encoded, if at all, is simple. Roughly speaking, qualities like being brilliant or very smart can be conceived of as acquired capacities, as traits that are innately possessed, or as a combination of both (Dweck, 2000, 2006). People will likely disagree about the relative importance of each, and about whether there are substantial differences across social groups. Despite those disagreements, most would accept inferences like the following: *if* Mary is less naturally brilliant than Susan, and all else is equal, Mary will have to work somewhat harder than Susan to achieve the same level of intellectual success. In this scenario, *smart* depends more on *hard work* for Mary than for Susan. The point is just that differences in the interdependency of *smart* specifically on *hardworking* is one way in which the brilliance-gender stereotype could be encoded in our target concepts.

Following this reasoning, Study 3 was designed to examine possible differences in the interdependencies between *hardworking* and *smart* in FEMALE vs MALE PROFESSORS. According to our hypothesis, *smart* should depend more on *hardworking* in FEMALE than in MALE PROFESSORS. To determine this, however, we cannot just ask participants how 'hardworking', say, a particularly accomplished female and male professor is, and then directly compare the average estimates for each gender. The reason is connected with the main limitation of Study 2: even if women are judged as more hardworking, it might be because of non-intellectual obstacles. To get around this obstacle, we opted for the following design. We described particularly successful female and male professors, and asked participants to estimate how hardworking they were, in terms of hours per week. We also included an additional feature that these individuals are thought, by their colleagues, to possess. In the control condition, this additional feature was 'being open-minded', and in the target condition it was 'bering very smart'.

If *smart* depends more on *hard work* for female than for male professors, then there should be greater difference, between the control and target conditions, in the estimates of hours per week of work for female than for male professors. Since the individuals described in all conditions are successful professors, any non-intellectual obstacles which are thought to specifically affect women will be reflected in the control condition, and will not influence the difference between that target and control condition, which is the value of interest. Our prediction is that there should be an interaction between *gender* and *feature*, such that the female and *smart* condition should have a stronger positive effect on hours of work relative to the female and control condition, than the effect of the male and *smart* condition relative to the male and control condition.

5.1 Methods

200 participants were recruited from Amazon's Mechanical Turk. Participants were randomly assigned to four conditions: *female&smart* ($N = 52$), *female&control* ($N = 48$), *male-&smart* ($N = 50$), and *male&control* ($N = 50$). The statements for each of the four conditions read:

> (Smart condition) Mary/Jack has recently become a professor at a prestigious university. Her/His colleagues think of her/him as a very smart person.
>
> (Control condition) Mary/Jack has recently become a professor at a prestigious university. Her/His colleagues think of her/him as a very open-minded person.

After being presented with one of the vignettes, we asked them to answer the following question:

(6) To get where s/he is now, how many hours a week did Mary/Jack work during the last few years? Please give your best estimate.

Participants were asked to rate the number of hours of work per week on a scale from 0 to 100 hours.

5.2 Results

The mean ratings for number of hours worked per week in the open-mindedness-control condition were lower ($M = 51.65, SD = 12.07$) than those obtained in the smart-condition ($M = 54.39, SD = 14.16$). This is as expected and provides evidence that the task performed by participants was the intended one. A small difference was observed between participants' ratings of the female protagonist ($M = 53.33, SD = 13.56$) and the male protagonist ($M = 52.77, SD = 12.94$). The average responses for each of the four conditions are presented in Table 4. A 2 X 2 ANOVA was conducted with *Gender* (*female, male*) and *Feature* (*smart, open-mindedness*) as independent factors, and *Amount of hours* as dependent measure. *Gender*, $F(1, 196) = 0.05; p < 0.818, \eta^2 < 0.01$ and

Feature, $F(1, 196) = 2.17, p = 0.142, \eta^2 = 0.01$, were not significant. Importantly, however, the interaction between *Gender* and *Feature* was significant, $F(1, 196) = 4.36; p = 0.038, \eta^2 = 0.02$.

	Male	Female
smart Prof.	52.20	56.50
open minded Prof.	53.34	49.89

Table 5: Results of Study 5: Mean values for male and female professors in terms of hours per week when either smartness or open-mindedness was emphasised.

5.3 Discussion

Study 3 examined whether there is a gender difference, in the prototypes for professors, in the dependency of *smart* on *hard work*. We asked participants to indicate how hardworking—in terms of hours per week—a female and male professor would have to be. The control and target conditions involved successful professors at prestigious universities, but only the target condition emphasised the feature of being especially smart. The results confirm our prediction: there was an interaction between gender and feature. Specifically, female professors whose smartness was emphasised were judged to have to work more hours, relative to their control condition, whereas male professors whose smartness was emphasised were *not* judged to have to work more hours, relative to their control condition. These results undermine the competing account raised in response to Study 2: namely, that female professors are conceived as more hardworking than male professors because they have to overcome additional obstacles that are not connected with presumed gender differences in raw or innate brilliance. For that view predicts that successful female professors at prestigious places would have to encounter these obstacles, regardless of whether they are described as in the target or control condition. It follows that this view cannot account for the differences in amount of work observed across conditions for female professors. Furthermore, if assumed additional obstacles are the explanation for why female professors are thought to be more hardworking, then participants should judge that female professors are more hardworking than male professors across the control conditions. But this is not the observed result: in contrast to the target condition, which emphasised their brilliance, in the control condition male professors are rated as more *hardworking* than female professors.[8] To sum up, Study 3 supports the view that *smart* depends more on *hardworking* in the prototype for FEMALE vs. MALE PROFESSOR. This directly supports Leslie et al. (2015)'s hypothesis that there is a brilliance-gender stereotype, which in this case

[8]The hypothesis that (people think that) women simply encounter more obstacles can perhaps be modified to explain the results, e.g. by arguing that open-mindedness causally interferes with the perception of female professors as hardworking. However, this modification is based on an ad hoc assumption, and it is not at all clear why being open-minded would interfere with the perception of female professors as hardworking.

is manifested in the prototypes for professors. In addition, it supports our main contention, namely, that the brilliance-gender bias is encoded in the dependency networks represented by the prototypes for FEMALE and MALE PROFESSOR.

6 General Discussion

Our Studies were designed to examine whether the brilliance-gender stereotype is encoded in the dependency networks of our prototypes for FEMALE and MALE PROFESSORS. The Preliminary Study showed that the key gender difference concerns the higher weight assigned to *hard work* in FEMALE PROFESSOR. Studies 1a-b suggest that this effect is not due to differences in the estimates of how likely it is that female vs. male professors are hardworking. Importantly, neither the Preliminary Study nor Studies 1a-b resulted in gender differences with respect to *smart*. In Studies 2-3, we then explored the role of *smart* and *hardworking* in the dependency networks of our target prototypes. Using a simple causal reasoning task, Study 2 showed that *hard work* is more central for FEMALE than for MALE PROFESSORS. This means that more features of FEMALE PROFESSOR depend on *hard work* than of MALE PROFESSOR. Although, as confirmed in our Preliminary Study, most features produced for the class of professors had to do with intellectual and mental qualities, Study 2 is still compatible with the possibility that there is no brilliance-gender stereotype and that the gender difference in the centrality of *hard work* is not due to dependency relations to features such as *smart*. Accordingly, in Study 3 we examined and confirmed that there are gender differences in the dependency of *smart* specifically on *hardworking*. Overall, our Studies support the hypothesis that the brilliance-gender stereotype is encoded in dependency networks of FEMALE and MALE PROFESSORS.

We should be clear about what we think we have and haven't achieved. We think we have key evidence for the view that notions such as centrality and dependency are required to fully understand how the brilliance-gender stereotype is encoded. We are not claiming, however, that this amounts to a complete picture of the brilliance-gender stereotype. Many important questions remain open. In particular, we need to carry out additional studies using other measures of both salient-statistical associations and centrality/dependency relations (see Sloman et al., 1998; Keil, 1989). These additional studies should allow us to refine some coarse assumptions that we made. For example, we have proceeded as if, in our basic-level concepts, features like *smart* and *brilliant* stand for one trait, which we can have to different degrees. Needless to say, there are likely subtle distinctions there. Accordingly, we should also explore whether we represent, in our concepts of social groups, different kinds/ways of being *smart*—e.g., in relation to quickness, creativity, or type of problem-solving capacity—and how this informs the brilliance-gender stereotype, including the special role of *hard work*. Although there are many open questions, we hope we have presented a serious case for our main contention, namely, that consequential biases are encoded in the dependency networks that we use to represent social groups. If, in

our empirical and philosophical studies of social cognition, we continue thinking of stereotypes as bundles of salient-statistical associations, we will miss this important dimension of the human mind. In the remainder of this General Discussion, we outline two implications of our account, both of which shed light on the uniqueness and importance of the notion of conceptual centrality for a more complete picture of bias in social cognition.

6.1 Compositionality, centrality, and cross-contextual stability

There is a key difference in the cross-contextual behaviour of salient-statistical vs. centrally encoded biases. Namely, biases which are encoded just in salient-statistical associations are less stable across contexts than those which depend on central features. To see why, we need to first briefly consider what might initially seem like an unrelated topic, namely, the behaviour of features in conceptual combination.

Philosophers have pointed out, and empirical studies have largely confirmed, that features which are associated with concepts merely via salient-statistical relations often do not survive combinatorial operations (Barsalou, 1987; Fodor, 1998; Fodor and Lepore, 2002; Hampton, 2006; Murphy, 2002; Rey, 1983). Suppose that MANE is a feature of the prototype LION, which 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 lion by their manes). Still, MANE does not survive some trivial conceptual combinations involving LION: consider, e.g., BABY LION, FEMALE LION and, with a bit of imagination, TRIMMED LION. These combinations are not 'special'; rather, they are simple interactive combinations, with the result that we move from basic level categories to more specific subcategories. In contrast, it is widely recognised that features that are central are more likely to survive similar conceptual combinations (Hampton, 1987, 2006; Murphy, 2002). To illustrate, take the feature BORN OF LION PARENTS, which is highly central for LION (cf. Keil, 1989), and consider your intuitions for the complex concepts BABY LION, FEMALE LION, and TRIMMED LION. Clearly, they all effortlessly inherit the feature BORN OF LION PARENTS. Importantly, the combinatorial properties of features affect the content and structure of many of the prototypes which we use in daily life. We often find ourselves in environments where we need to sub-categorise. For example, suppose you are at a lion nursery. To function in that environment, it is important that you operate with the representation BABY LION. In this way, you will not be looking out for manes or constantly panic and perhaps hide in some closet; however, you will still assume that the baby lions were born in the usual way.

At this point, we can see why the degree of centrality of the features which encode a bias is an important determinant of the bias' wider role in social cognition. Suppose we have shown that feature f is more central to our conception of FEMALE than to our conception of MALE. This means that f will have a greater degree of cross-contextual stability for FEMALE; specifically, f will be more likely to survive conceptual combinations and sub-categorisation involv-

ing FEMALE than those involving MALE. For example, f will be more likely to survive composition into subcategories such as FEMALE LAWYER, FEMALE DOCTOR and FEMALE POLITICIAN than to survive into MALE LAWYER, MALE DOCTOR and MALE POLITICIAN. Furthermore, information about the details of the dependency networks allow us to make even more refined predictions. To illustrate, if, as we argued, SMART depends more on HARD WORK for FEMALE than for MALE PROFESSORS, then combinations and sub-categorisations that tinker with HARD WORK should have a stronger effect on SMART in the case of FEMALE PROFESSORS. For example, we have seen that participants think of professors as smart. Suppose we consider the class of LAZY MALE/FEMALE PROFESSORS. These combinations lower the rating of HARD WORK compared to the default ratings it gets in PROFESSORS. Given the gender differences in the dependency networks, this is predicted to more negatively affect the perceived degree of SMARTNESS for FEMALE than for MALE PROFESSORS.

To sum up, notions like conceptual centrality and dependency networks are crucial to understanding the wider role of biases in social cognition. In particular, they are crucial to determining the cross-contextual stability of target biases, including biases encoded in salient-statistical associations. By incorporating these notions into our accounts of stereotypes and prototypes, we also open up a very rich set of questions and predictions that we can empirically examine.

6.2 The varieties of implicit bias

While theorists have typically interpreted implicit biases as 'mere associations' between groups and salient or typical traits, some philosophers (Mandelbaum, 2016; Levy, 2015) and social psychologists (De Houwer, 2014) have argued that implicit biases must have a richer conceptual or propositional structure. Broadly speaking, we agree that, for the purposes of understanding how implicit biases inform social cognition and discrimination, focusing solely on mere group-trait associations is misguided. However, we note two key ways in which our approach differs from existing views.

First, while these theorists have argued that implicit biases must have some richer conceptual structure, they haven't, at this point, said much about precisely what this structure might be. Our research constitutes a significant step forward in that we provide empirical evidence for at least one concrete way in which implicit biases can be encoded in conceptual structures (beyond salient-statistical associations), namely, in dependency networks.

Second, the debate between 'associative' and 'propositional' interpretations of implicit bias has so far centered on the nature of the psychological constructs that explain timed measures of response latency or error, such as the IAT. By contrast, we have argued that important forms of bias may fail to be detected by such measures altogether. That is, we do not predict that dependency networks will generally correlate with IAT results. To the contrary, we predict that the features that are (perceived to be) salient or typical of certain groups will often differ from those that are (perceived to be) central to those groups (see

e.g., Sloman et al., 1998). On our view, both salient-statistical associations, some of which are manifest in IAT scores, and dependency networks, which are manifest in ways we have explored here, play a role in social cognition and discrimination. Salient-statistical associations and central features thus need not be seen as 'competing' to explain the same set of phenomena, but may instead explain different phenomena, or make different contributions to the explanation of phenomena. In fact, we find the assumption that there is 'one' sort of bias—whether associative or propositional—driving all discrimination to be implausible on its face and empirically unsupported. We hope to have made some headway towards appreciating the diverse ways in which social cognition can encode biases and produce discriminatory outcomes.

References

Banaji, M. R. and A. G. Greenwald (2013). *Blindspot: The hidden biases of good people*. New York: Delacorte Press.

Barsalou, L. 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*, pp. 101–140. Cambridge, UK: Cambridge University Press.

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

Berinsky, A. J., G. A. Huber, and G. S. Linz (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis 20*, 351–368.

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

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

Di Bella, L., E. Miles, and J. Saul (2016). Philosophers explicitly associate philosophy with maleness: An examination of implicit and explicit gender stereotypes in philosophy. In J. Saul and M. Brownstein (Eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, pp. 283—308. Oxford: Oxford University Press.

Dweck, C. S. (2000). *Self-theories: their role in motivation, personality and success*. Philadelphia, PA: Psychology Press.

Dweck, C. S. (2006). *Mindset: the new psychology of success*. New York: Random House.

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

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Fodor, J. and E. Lepore (2002). *The Compositionality Papers*. Oxford: Oxford University Press.

Gelman, S. A. and H. W. Wellman (1991). Insides and essences: early understanding of the non-obvious. *Cognition 38*, 213–244.

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

Greenwald, A. G., T. A. Poehlman, E. L. Uhlmann, and M. 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.

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

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

Johnson, C. and F. C. Keil (2000). Explanatory knowledge and conceptual combination. In F. C. Keil and R. A. Wilson (Eds.), *Explanation and Cognition*. The MIT Press.

Keil, F. C. (1989). *Concepts, Kinds and Cognitive Development*. Cambridge, MA: The MIT Press.

Keil, F. C. (2003). Categorization, causation, and the limits of understanding. *Language and Cognitive Processes 100*(2), 663–692.

Keil, F. C. (2006, Jan). Explanation and understanding. *Annual Review of Psychology 57*(1), 227–254.

Lane, K. A., M. R. Banaji, B. A. Nosek, and A. G. Greenwald (2007). Understanding and using the implicit association test: Iv. In B. Wittebrink and N. Schwarz (Eds.), *Implicit measures of attitudes*, Chapter 3, pp. 59–102. New York: Guilford.

Leslie, S.-J., A. Cimpian, M. Meyer, and E. 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. (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.

McRae, K., G. S. Cree, M. S. Seindenberg, and C. McNorgan (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral reaserch methods 37*(4), 547–559.

Meyer, M., A. Cimpian, and S.-J. Leslie (2015). Women are underrepresented in fields where success is believed to require brilliance. *Frontiers in Psychology 6*(235), 1–12.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: The MIT PressMIT Press.

Paolacci, J. and J. Chandler (2014). Inside the turk: understanding mechanical turk as a participant pool. *Current Directions in Psychological Science 23*, 184–188.

Pinker, S. (2007). *The Stuff of Thought: Language as a Window into Human Nature*. New York: the Penguin Group.

Prinz, J. (2002). *Furnishing the Mind*. Cambridge, MA: MIT Press.

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

Rosch, E. (1999). Principles of categorization. In E. Margolis and S. Laurence (Eds.), *Concepts: Core Readings*, Chapter 8, pp. 189–206. Cambridge, MA: The MIT Press.

Rosch, E. (2011). "slow lettuce": Categories, concepts, fuzzy sets, and logical deduction. In R. Belohlavek and G. J. Klir (Eds.), *Concepts and Fuzzy Logic*, Chapter 4, pp. 89–120. Cambridge, MA: The MIT Press.

Sloman, S. A., B. C. Love, and W.-K. Ahn (1998). Feature centrality and conceptual coherence. *Cognitive Science 22*(2), 189–228.

Valian, V. (1998). *Why so slow? The advancement of women*. Cambridge, MA: The MIT Press.