Introduction to the Philosophy and Cognitive Science of Stereotypes and Social Biases

PHIL 482, WN 2019, Tu 4:00 - 7:00 PM, Room: G026 TISCH

Instructor: Guillermo Del Pinal, Weiser Hall 914, delpinal@umich.edu

Office Hours: Weiser 926, Mondays 10-12pm and by appointment.

Prerequisites: None

Course Outline:

Concepts, prototypes, and stereotypes	2 weeks
Philosophical reflections on implicit bias	3 weeks
Concepts, dependencies, and centrality	2 weeks
Central biases and gender stereotypes	2 weeks
Generics and cognitive biases	2 weeks
Dual character concepts	1 week
Machine Learning and human-like biases in AI	2 week

Course Description: Our mental representations of social categories often encode biases which can deeply affect our social judgments and behavior. In this class, we will explore the ways in which social biases can be encoded in our representations of social groups ('stereotypes/concepts'), with the goal of trying to understand their impact on social judgments and behavior. One influential tradition in social psychology models 'stereotypes' as bundles of salient-statistical features which we use to represent categories (e.g., 'striped' is a typical and salient feature for 'tigers'). This approach has led to important advances in the development of experimental measures, specific biases discovered, and in our understanding of the effect of stereotypes on social cognition. However, this approach also has substantial limitations. Amongst philosophers and cognitive scientists, it is increasingly recognized that concepts and stereotypes have a much richer structure than traditionally assumed: e.g., they encode information about the causal dependencies between features (e.g., that 'cars' have a 'steering wheel' so that they can be 'controlled') and the degree of centrality of a features for a category (e.g., that 'used for driving' is more central for 'cars' than 'being made in Detroit'). In this class, we will examine recent work in philosophy and cognitive science that approaches the study of social biases by drawing on these richer accounts of the structure of human concepts. Building on this work, we will try to (i) improve our taxonomy of and measures for socially significant biases, (ii) understand the relation between specific types of biases and their unique effects on social judgments, and (iii) explore the consequences of specific kinds of biases on questions about individual responsibility and strategies for effective interventions.

Materials: All readings and assignments will be posted on our U-M Canvas course webpage

Grading: [15%] class participation + 5 critical questions on (required) readings (≈100-250 words), [40%] one presentation (≈ 20 mins + 10 mins for discussion), [45%] one term paper (≈ 3,000 words).

Academic Integrity: Each student in this course is expected to abide by the University of Michigan's Honor Code.

Accommodations for students with disabilities: In compliance with the University of Michigan policy and equal access laws, I am available to discuss appropriate academic accommodations that may be required for student with disabilities. Students are encouraged to register with Services for Students with Disabilities to verify their eligibility for appropriate accommodations.

Inclusivity statement: Our members represent a rich variety of backgrounds and perspectives. The Philosophy Department and The Weinberg Institute for Cognitive Science are committed to providing an atmosphere for learning that respects diversity. While working together to build this community we ask all members to: (i) share their unique experiences, values and beliefs, (ii) be open to the views of others, (iii) honor the uniqueness of their colleagues, (iv) appreciate the opportunity that we have to learn from each other in this community, (v) value each other's opinions and communicate in a respectful manner, (vi) keep confidential discussions that the community has of a personal (or professional) nature, (vii) use this opportunity together to discuss ways in which we can create an inclusive environment in this course and across the University of Michigan community.

Extra Help: Please do not hesitate to come to my office during office hours or make an appointment to discuss any aspect of the course.

University Attendance Policy: Students are expected to attend classes regularly. A student who incurs an excessive number of absences may be withdrawn from a class at the discretion of the instructor.

Tentative reading schedule: (**Last Update: April 22, 2019**)
Key for readings: R = Required, O = Optional

Week 1 (1/15): Concepts, prototypes, and stereotypes: Foundations

(R) Rosch (1978). Principles of categorization.
(O) Wittgenstein (1953). Selections from *The Philosophical Investigations*: §65 -§78

***Notes:** An excellent background/reference source for weeks 1 to 4 of the course is the Stanford Encyclopedia of Philosophy entry 'Implicit Bias' (https://plato.stanford.edu/entries/implicit-bias/).

Week 2 (1/22): Concepts and stereotypes: Implicit association test and social biases

(R) Nosek et al. (2011) Implicit social cognition. *Trends in Cognitive Science.*
(O) Banaji and Greenwald (2013). Selections from *Blindspot* (Ch. 5-6)
(O) Greenwald et al. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology.*

***Notes:** To get a sense for how the Implicit Association Test (IAT) works, I highly recommend that you take at least one test at Project Implicit (https://implicit.harvard.edu/implicit/).

Week 3 (2/5) (NO class on 1/29): Philosophical reflections: Beliefs and implicit bias

(R) Gendler (2008). Alief and belief. *Journal of Philosophy.*
(R) Gendler (2008). Alief in action and reaction. *Mind & Language.*
(O) Mandelbaum (2014), Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous.*
(O) Madva (2015). Why implicit attitudes are probably not beliefs. *Synthese.*
(O) Levy (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous.*

***Notes:** For background, read §2-§3 of 'Implicit Bias' (https://plato.stanford.edu/entries/implicit-bias/ (SEP). Our student presentations for this week are: Jimmy A Kenny (presenting Mandelbaum 2014) and Elizabeth McGowan (presenting Levy 2015).

Week 4 (2/12): Philosophical reflections: Responsibility and implicit bias

(O) Antony, L .(2016), Bias: Friend or Foe? Reflections on Saulish skepticism. in *Implicit Bias and Philosophy (v1)*
(O) Brownstein, M. (2015). Attributionism and Moral Responsibility for Implicit Bias. *Review of Philosophy and Psychology*
(O) Zheng, R. (2016). Attributability, Accountability, and Implicit Bias. *Implicit Bias and Philosophy (v2).*
(O) Madva, A. (2017). Implicit bias, moods, and moral responsibility. *Pacific Philosophical Quarterly*

***Notes:** Flash talks day! For background, read §4 of 'Implicit Bias' (https://plato.stanford.edu/entries/implicit-bias/ (SEP). Student presentations for this week are: Adam Bean (on Brownstein 2015), Brendan Duff (on Antony 2016), and David Kamper (on Zheng 2016).

Week 5 (2/19): Skepticism about Implicit Bias: Criticisms of the Implicit Association Test

(R) Singal, J. Psychology's favorite tool for measuring racism isn't up to the job. (link: https://www.thecut.com/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html)
(R) Brownstein, M., Madva, A. and Gawronski, B. (2019, ms). Understanding Implicit Bias: How the Critics Miss the Point.

***Notes:** Guest talk by Prof. Chandra Sripada (University of Michigan, Departments of Psychiatry and Philosophy)!

Week 6 (2/26): Concepts and feature dependencies: Foundations

(R) Sloman et al. (1998). Feature centrality and conceptual coherence. *Cognitive Science.*
(O) Sloman (2014). Two systems of reasoning: An update.
(O) Danks, D. (2014). Concepts, categories and inference. Ch. 5 of *Unifying the Mind.*

***Notes:**

Week 7 (3/12) (3/5 is Spring Break): Concepts and feature dependencies: Essentialism

(R) Gelman (2004). Psychological essentialism in children. *Trends in Cognitive Sciences.*

(R) Leslie (2013). Essence and natural kinds: When science meets preschooler intuition. *Oxford Studies in Epistemology.*
(O) Salomon and Cimpian (2014). The inherence heuristic as a source of essentialist thought. *Personality and social psychology bulletin.*
(O) Haslam et al. (2000), Essentialist beliefs about social categories. *British Journal of Social Psychology*

***Notes:** Student presentations for this week are: Larisa Kokubo (on Haslam et al 2000)

Week 8 (3/19): Gender and race stereotypes and biases

(O) Leslie et al. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science.*
(O) Bian et al. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science.*
(O) Mandalaywala, T. M. and Ranger-Murdock, G. and Amodio, D. M. and Rhodes, M. (2018). The Nature and Consequences of Essentialist Beliefs About Race in Early Childhood. *Child Development*
(O) Ho, A. K., Roberts, S. O., & Gelman, S. A. (2015). Essentialism and racial bias jointly contribute to the categorization of multiracial individuals. *Psychological Science.*

***Notes:** Student presentations for this week: Meena Seetharaman (on Leslie et al. 2015), Annie Rodgers (on Bian 2017), Mansi M Brahmbhatt (on Mandalaywala et al 2018).

Week 9 (3/26): Stereotypes, 'explanations', and central biases

(R) Del Pinal and Spaulding (2018). Conceptual centrality and implicit bias. *Mind & Language.*
(O) Del Pinal, Madva and Reuter (2017). Stereotypes, conceptual centrality and gender bias. *Ratio.*

***Notes:**

Week 10 (4/2): Generics and default generalizations

(R) Leslie (2018). The Original Sin of cognition: Fear, prejudice and generalization. *The Journal of Philosophy.*
(O) Leslie, S. J. (2008). Generics: Cognition and acquisition. *Philosophical Review*
(O) Prasada et al. (2013). Conceptual distinctions amongst generics. *Cognition.*

***Notes:** Student presentations for this week: Anthony Bryant (on Leslie 2018), Michael Villarica (on Prasada et al. 2013)

Week 11 (4/9): Generics, stereotypes and over-generalization

(R) Jussim (2017). Precis of *Social Perception and Social Reality.* Focus on §1 and §8-10.
(R) Bian and Cimpian (2017). Are stereotypes accurate? *Behavioral and Brain Sciences* 22.
(O) Hammond and Cimpian (2017). Investigating the cognitive structure of stereotypes: Generic beliefs about groups predict social judgments better than statistical beliefs. *Journal of Experimental Psychology.*
(O) Cimpian et al. (2010). Generic beliefs require little evidence for acceptance but have powerful implications. *Cognitive Science.*

(O) Khemlani et al. (2012). Inferences about members of kinds: The generics hypothesis. *Language and Cognitive Processes* 27(6): 887-900

(O) Tasimi, Gelman, Cimpian and Knobe (2016). Differences in the evaluation of generic statements about human and non-human categories. *Cognitive Science.*

***Notes:** Student presentations for this week: Elle Konrad (on Cimpian et al 2010), Katie Wagner (on Khelmani et al. 2012), Laurita Mansour (on Tasimi et al. 2016).

Week 12 (4/16): Dual Character concepts, social roles and the normative dimension

(R) Knobe et al. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition.*

(O) Leslie (2015). "Hillary Clinton is the only man in the Obama Administration": Dual character concepts, generics and generalization. *Analytic Philosophy.*

(O) Del Pinal and Reuter (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science.*

(O) Strohminger, N. et al. (2017) The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science.*

***Notes:** Student presentations for this week: Sarah Marks (on Leslie 2015), Sydney Angel (on Strohminger et al. 2017)

Week 13 (4/23): Lessons from AI: Machine learning and social biases

(R) Caliskan et al. (2017). Semantics derived automatically from language corpora contain human like biases. *Science.*

(O) Greenwald, A. (2017). An AI stereotype catcher. *Science*

(O) Garg et al. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences.* vol. 115(16):

***Notes**: Presentations for this week: Raphael Korosso (on Garg et al. 2008), Victoria Johnson (on Caliskan et al. 2017)

Week 14 (No Class!): Lessons from AI: Machine learning, social biases and interventions

(R) Zou, J. and Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature*

(O) Bolukbasi et al. (2016). Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings. *29th Conference on Neural Information Processing Systems (NIPS).*

(O) Barocas and Selbst (2016). Big data's disparate impact. *California Law Review*, vol. 104: 671-729.

***Notes:** Although there is no class, I left readings up in case you want to follow up on your own on the topics covered in week 13.