Graduate Philosophy of Psychology: Social Bias Syllabus

Carolina Flores

Class Hours: W 9:30-12:30
Classroom: Crown 201
Office: Cowell College Faculty Office Addition, Office 104
Office Hours: W 14:30-15:30; other times by appointment
Email: caro.flores@ucsc.edu

Required Texts

All readings, handouts, assignments, and announcements will be posted on Canvas.

Course Description

Sexism, racism, and their ilk are everywhere. Among other places, they are in our minds, through biases about social groups, and in algorithms and our informational environments. But what exactly are biases and in what ways are they implemented in us and in machines? What can we learn about the structure of thought, and about human nature, by studying social biases? And what can we do to reduce and ameliorate them?

In this class, we will explore how biases are encoded in human representations of social categories, and how they impact social judgments, attention, perception, and behavior. We will also investigate how social biases are reproduced by algorithms, further entrenching structural injustices. To do so, we will examine both well-established and cutting-edge psychological and philosophical models of stereotypes and bias in cognitive science and discuss their limits. Against this background, we will then investigate social biases hidden in data and algorithms generated via machine learning. Finally, we will consider ethical questions all of this raises for us as a society.

Course Goals

In this course, you will:

- Acquire knowledge of central debates in empirically informed philosophy about the nature and mechanisms of social bias, both in humans and in machines.
- Come to grasp central concepts, distinctions, and theories in the study of social bias and of cognitive architecture.
- Develop the ability to assess philosophical accounts of social bias in light of empirical findings.
- Develop the ability to sketch interventions aiming to address instances of social bias based on our best knowledge of the phenomena.

In pursuing these course-specific goals, you will also acquire the following general skills:

- To engage in close and charitable readings of sophisticated arguments.
- To criticize views by giving focused objections to them and anticipating replies.
- To communicate complex ideas effectively and concisely in your writing.

- To engage in respectful, reasoned, and passionate debate with peers about complex topics that lack clear answers, and to use such debate as a tool for understanding.

Schedule of Topics and Readings *(subject to change)*

Date	Topic	Readings
Unit 1. Social Bias in Humans.		
Oct 4	Introduction. Implicit bias and the IAT.	• Take two of the IAT tests at Project Implicit • Mahzarin Banaji and Anthony Greenwald (2013), *Blindspot*, Chapters 5-6. • Brian Nosek et al. (2011), Implicit social cognition: From measures to mechanisms.
Oct 11	Skepticism about implicit bias research.	• Jesse Singal, Psychology's favorite tool for measuring racism isn't up to the job. • What can we learn from the Implicit Association Test? A Brains Blog Roundtable. • Chandra Sripada (2022), Whether implicit attitudes exist is one thing, and whether we can measure them effectively is another.
Oct 18	Bias and essentialism about social groups.	• Sarah-Jane Leslie (2017), The original sin of cognition. • Elli Neufeld (2022), Psychological essentialism and the structure of concepts. • Rebecca Peretz-Lange (2021), Why does social essentialism sometimes promote, and other times mitigate, prejudice development? A causal discounting perspective. **APPLIED ROUTE: CHOICE OF TOPIC AND MINIMAL READING LIST DUE**
Oct 25	Bias in our concepts of social groups.	• Guillermo Del Pinal and Shannon Spaulding (2018), Conceptual centrality and implicit bias. • Guillermo Del Pinal, Alex Madva, and Kevin Reuter (2017), Stereotypes, conceptual centrality, and gender bias. • Brains Blog Symposium on Del Pinal and Spaulding, Conceptual centrality and implicit bias.
Nov 1	Bias beyond cognition: perception and attention.	• B.K. Payne (2001), Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. • Celine Leboeuf (2020), The embodied biased mind. • Jessie Munton (2021), Prejudice as the misattribution of salience. **APPLIED ROUTE: ANNOTATED BIBLIOGRAPHY DUE**
Unit 2. Social Bias in Machines.		
Nov 8	Algorithmic bias.	• Sina Fazelpour and David Danks (2022), Algorithmic bias: senses, sources, solutions. • Gabby Johnson (2021), Algorithmic bias: On the implicit biases of social technologies.

		• Lily Hu (2021), What is "race" in algorithmic discrimination on the basis of race? **TRADITIONAL ROUTE: SHORT PAPER 1 DUE**
Nov 15	Bias in word embeddings.	• Background: Ben Levinstein (2023), A conceptual guide to transformers (all five parts) • Aylin Caliskan et al. (2022), Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. • Melissa Heikkila (2023), AI language models are rife with different political bias • Open AI cofounder responds to Elon Musk's criticism that ChatGPT is too 'woke': 'We made a mistake' **APPLIED ROUTE: LIST OF POTENTIAL INTERVENTIONS DUE**
Unit 3. Social Bias: Problems and Solutions.		
Nov 22	Social bias and accuracy.	• Louise Antony (2016), Bias: Friend or foe? • Lee Jussim et al. (2015), Stereotype (in)accuracy in perceptions of groups and individuals • Lin Bian and Andrei Cimpian (2017), Are stereotypes accurate? A perspective from the cognitive science of concepts. • Jessie Munton (2019), Beyond accuracy: Epistemic flaws with statistical generalizations.
Nov 29	Algorithmic bias and fairness.	• Brian Hedden (2021), On statistical criteria of algorithmic fairness. • Clinton Castro (2022), Just machines. • Kathleen Creel and Deborah Hellman (2022), The algorithmic Leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems.
Dec 6	Conclusion + Presentations.	No reading; presentations of final papers and projects.
Dec 15	**FINAL PAPERS/EXECUTIVE SUMMARIES DUE**	

Course Requirements

- Attendance and participation (10%)
- Weekly forum posts (20%)

For the remaining 70%:

Traditional route

- Two short papers (around 3,000 words), each worth 35% of the grade OR
- One short paper (around 3,000 words) expanded into a longer paper (around 7,000 words) after comments, for a total of 70% of the grade awarded to the final paper

OR

Applied route

Pair work: developing an intervention to address a particular instance of one kind of social bias discussed in the class. You should identify a particular kind of bias in a particular context (e.g. implicit gender bias in instructors in university classrooms; algorithmic racial bias in financial aid algorithms). This will include the following steps:

- Oct 18: Choice of topic and minimal reading list (5%)
- Nov 1: 3/4-page Annotated bibliography (5%)
- Nov 15: List of 5 potential interventions, including at least 2 wild ones (5%)
- Dec 6: Presentation/pitch of the intervention (30%)
- Dec 15: 2,000-word executive summary of the intervention. This should include identifying the problem, a brief overview of the existing research and interventions, and a description of the proposed intervention, its intended goals, audience, scope, and potential risks (25%)

Attendance and Participation

Attending every class is compulsory. Please notify me in advance if you expect to miss a class. Come prepared: do the reading in advance, pay attention, engage with your peers, ask (clarificatory or substantive) questions, and contribute to our collective understanding.

Forum Posts

Once a week, everyone will post a comment to that week's Canvas forum. This can be an independent post or a reply to someone else's post. It should be a **paragraph** engaging with some of the reading for that week. This can be

a) a focused question on the reading (for example, a request for clarification on what view the author is putting forward, a question about how an argument works, or a question on how the view connects to other claims we have discussed in class),
b) an objection to a specific claim (that you identify by including in your post) made in the reading,
c) an additional argument for the view offered in the reading, or
d) an example that supports or causes trouble for a claim made in the reading.

Comments are due Tuesday at 2pm. Partial credit will be awarded for late responses. You can miss or drop one post over the course of the entire quarter without penalty.

Papers

Papers should be submitted as .pdfs. Except in extreme conditions, extensions must be granted well before the due date; late papers will be downgraded 1/3 grade per day. The paper should be a reasoned defense of a view, addressing some of the topics discussed in the class.

I advise you to consult with me well advance of the deadline about the topic of your papers, and to attend office hours to discuss an outline and/or to brainstorm ideas. We will spend time in class talking about how to write a good philosophy paper.

A note on ChatGPT and other AI tools

Using ChatGPT to write a paper for you is plagiarism and a waste of your time as a participant in this course, and therefore not allowed in this class.

However, I allow reasonable uses of AI as an aid in research and writing. For example, you may use ChatGPT to help you rephrase or streamline your writing; to get ideas for additional arguments or counter-arguments to a view you are exploring; to clarify your understanding of concepts and theories discussed in the class. If you do so as an aid in writing, please write a brief explanation at the end of your paper of how and where you used ChatGPT. I advise caution, as ChatGPT has been know to produce non-existent sources or invalid arguments. If using ChatGPT to check your understanding or research sources, always double check.

Other Expectations

I expect you to be familiar with and to abide by the university's policy on academic and intellectual integrity. Violations of this policy include cheating, fabrication, plagiarism, denying others access to information or material, and facilitating violations of academic integrity.

I also expect all participants to observe basic norms of civility and respect. This means stating your own views directly and substantively: focusing on reasons, assumptions, and consequences rather than on who is offering them, or how. And it means engaging other's views in the same terms. No topic or claim is too obvious or controversial to be discussed; but claims and opinions have a place in the discussion only when they are presented in a respectful, collegial, and constructive way.

Accommodations

If you need to be absent for religious observances, let me know in advance. I will excuse without penalty students who are absent from class because of religious observance and allow the make-up of work missed because of such absence.

UC Santa Cruz is committed to creating an academic environment that supports its diverse student body. If you are a student with a disability who requires accommodations to achieve equal access in this course, please submit your Academic Access Letter from the Disability Resource Center (DRC) to me privately during my office hours or by appointment, as soon as possible in the academic quarter, preferably within 1 week. I also encourage you to discuss with me ways we can ensure your full participation in this course. I encourage all students who may benefit to learn about the DRC and the UCSC accommodation process. You can visit the DRC website at drc.ucsc.edu. You can make an appointment and meet in-person with a DRC staff member. The phone number is 831-459-2089, or email drc@ucsc.edu.