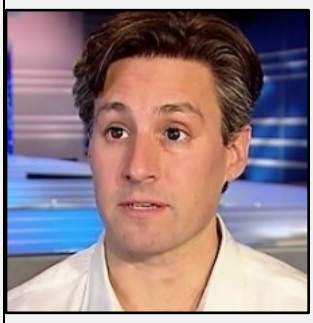
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

2.2.22



"We're Thinking About Data Privacy All Wrong "
Reid Blackman

Founder and CEO of Virtue Consultants
Chief Ethics Officer, Government Blockchain Association
Advisory Board, Ethics Grade, Hatch, & EY

Abstract: The received wisdom is that almost everyone's data privacy is violated every second of every day. Our data is meticulously collected, organized, and analyzed by everyone from tech behemoths like Google, Amazon, and Facebook to your local online retailer collect. The received wisdom, though well-intentioned, is misguided and counterproductive. It misconceives the nature of privacy and, what's worse, draws our attention away from what's really at issue: stopping the wrongs perpetuated by companies with effective regulation—focused on banning individually and socially treacherous behaviors.

Companies should be held to account when they create the grounds for the spread of democracy-undermining misinformation. They should be held to account for when their algorithms cause widespread depression among teenage girls. They should be held to account when they engage in literal surveillance, as when they deploy facial recognition software to identify particular people. These are the kinds of wrongs we need to keep our legal and regulatory eye on. Focusing on personal data privacy misses the point.

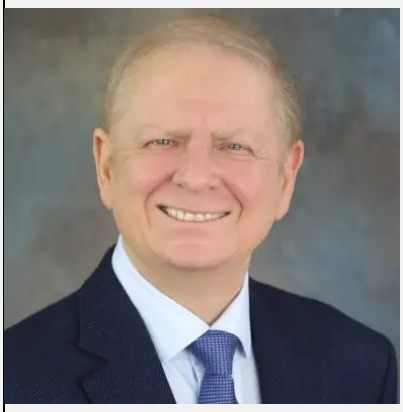
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

2.9.22



"Ethical Principles for Artificial Intelligence"

Paul Thagard

Distinguished Professor Emeritus of Philosophy,
The University of Waterloo
Fellow of the Royal Society of Canada, the Cognitive Science Society,
& the Association for Psychological Science

Abstract: The dramatic theoretical and practical progress of artificial intelligence in the past decade has raised serious concerns about its ethical consequences. In response, more than 80 organizations have proposed sets of principles for ethical artificial intelligence. The proposed principles overlap in their concern with values such as transparency, justice, fairness, human benefits, avoiding harm, responsibility, and privacy. This talk argues that the use of principles in medical ethics provides a good model for bringing order to the overabundance of AI proposals. In contrast to the hundreds of principles that have been proposed for AI ethics, a common approach to medical ethics gets by with four key principles concerning autonomy, justice, benefits, and avoiding harm. By sampling from current AI proposals, I infer that AI ethical principles fall under these four. Then the considerations that provide legitimacy to the four principles of medical ethics carry over to more specific AI principles.

Background Reading: "Medical Ethics Four Principles Plus Attention to Scope" Raanan Gillon

Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)
https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

2.16.22

"Legitimacy, Authority, and the Value of Explanations"

Seth Lazar

Professor of Philosophy,
Australian National University
Distinguished Research Fellow,
Oxford Institute for Ethics in AI

Abstract: As history's most potent corporations and the diminished neoliberal state meet rapid advances in Artificial Intelligence (AI), we are increasingly subject to power exercised by means of computational systems. Machine learning, big data, and related technologies now underpin vital government services from criminal justice to tax enforcement, public health to social services, immigration to defense. Our societies are increasingly dependent on systems whose operations are not being adequately explained to democratic citizens. A deeper normative analysis of the problem of opaque computational systems can shed light on the importance of explanations for the exercise of political power. A better understanding of why explanations matter can help us determine what kinds of explanations are owed, and to whom. To be morally permissible, new power relations must in general meet standards of procedural legitimacy and proper authority. Legitimacy and authority depend, in turn, on those who exercise power being able to explain their decisions to the community on whose behalf or by whose leave that power is exercised. The content of these explanations is determined by their function: to advance legitimacy and authority. They are owed, first and foremost, to the political community, not (on these grounds, at least) to the decision subject.

Background Readings: "Justification and Legitimacy" A.J. Simmons
"The Right to Explanation" Kate Vredenburgh

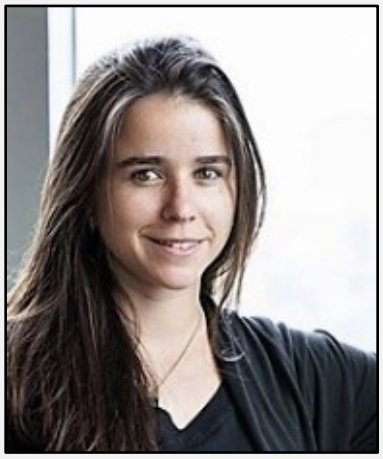
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

2.23.22



"A Case for Humans-in-the-Loop: Decisions in the Presence of Misestimated Algorithmic Scores"

Maria De-Arteaga

Assistant Professor of Information, Risk, and Operation Management,
Core faculty member, Machine Learning Laboratory,
University of Texas at Austin

Abstract: The increased use of algorithmic predictions in sensitive domains has been accompanied by both enthusiasm and concern. To understand the opportunities and risks of these technologies, it is key to study how experts alter their decisions when using such tools. In this paper, we study the adoption of an algorithmic tool used to assist child maltreatment hotline screening decisions. We focus on the question: Are humans capable of identifying cases in which the machine is wrong, and of over-riding those recommendations? We first show that humans do alter their behavior when the tool is deployed. Then, we show that humans are less likely to adhere to the machine's recommendation when the score displayed is an incorrect estimate of risk, even when overriding the recommendation requires supervisory approval. These results highlight the risks of full automation and the importance of designing decision pipelines that provide humans with autonomy

Background Reading: "Algorithm Aversion People Erroneously Avoid Algorithms after Seeing Them Err" Berkeley Dietvorst et al.

Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

3.2.22

"Fine-Grained Explanations Using Markov Logic"

Deepak Venugopal

Associate Professor of Computer Science, The University of Memphis
Affiliate, Institute for Intelligent Systems, The University of Memphis

Abstract: With the rapid growth in the use of AI for decision-making, there is an unquestionable need for Explainable Artificial Intelligence (XAI). Specifically, understanding why a decision is being made is as important as making accurate decisions. Indeed, there is an ongoing effort to pass legislation that requires decisions made by AI models to be explainable. Transparency of AI models will also significantly impact other critical fundamental issues in AI including fairness and bias. Over the last few years, XAI has made significant progress where several "black-box" explainers have been developed that explain predictions made by Machine Learning-based classifiers. In this talk, I will discuss XAI for relational data. I will motivate the use Markov Logic Networks (MLNs) as a language for XAI that can provide rich explanations that utilize relationships among instances.

Background Video: Unifying Logical and Statistical AI with Markov Logic

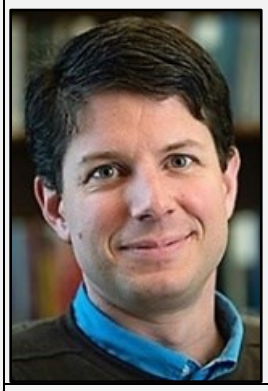
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

3.16.22



"Discovering Bias"
David Danks

Professor of Data Science & Philosophy, UC-San Diego
Affiliate Faculty, Department of Computer Science & Engineering, UC-San Diego

Abstract: Algorithms are increasingly shaping and informing human decision-making; in some cases, they are even making decisions that were previously the sole province of humans. Many recent incidents have led to harms from algorithmic bias: the ways that algorithms can embody, implement, maintain, and even create ethically, psychologically, and societally problematic biases. Several different measures of algorithmic bias have been proposed under the label of "fair machine learning," though these measures apply to algorithmic models more generally. However, current measures of bias and fairness focus solely on identifying surface signals, rather than discovering the sources, structures, and networks of bias that produce those patterns. In this talk, I will start to address the deeper problem of discovering the underlying (unobserved) societal, legal, psychological, and sociological biases and mechanisms that produce unjust (statistical) disparities. Existing approaches to algorithmic bias and fair ML regard it as a technical, statistical, and computational issue when it is actually a sociological, psychological, and philosophical one. I will show how to address (some of) the challenges of algorithmic bias in a fundamentally new way by reframing it as a discovery problem, rather than a measurement problem.

Background Readings: "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning" Sam Corbett-Davies and Sharad Goel
"Algorithmic bias: Senses, sources, solutions" Sina Fazelpour & David Danks

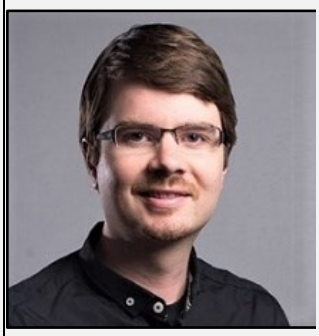
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

3.23.22



"Bias Preservation in Fair Machine Learning"

Brent Mittelstadt

Senior Research Fellow & British Academy Postdoctoral Fellow,
Oxford Internet Institute
Turing Fellow, Alan Turing Institute
Member, UK National Statistician's Data Ethics Advisory Committee

Abstract: Western societies are marked by diverse and extensive biases and inequality that are unavoidably embedded in the data used to train machine learning. Algorithms trained on biased data will, without intervention, produce biased outcomes and increase the inequality experienced by historically disadvantaged groups. Recognizing this problem, much work has emerged in recent years to test for bias in machine learning and AI systems using various fairness and bias metrics. Often these metrics address technical bias but ignore the underlying causes of inequality and take for granted the scope, significance, and ethical acceptability of existing inequalities. In this talk, I will introduce the concept of "bias preservation" to assess the compatibility of fairness metrics used in machine learning against the notions of formal and substantive equality. The fundamental aim of EU non-discrimination law is not only to prevent ongoing discrimination, but also to change society, policies, and practices to 'level the playing field' and achieve substantive rather than merely formal equality. Based on this, I will introduce a novel classification scheme for fairness metrics in machine learning based on how they handle pre-existing bias and thus align with the aims of substantive equality. Specifically, I will distinguish between 'bias preserving' and 'bias transforming' fairness metrics. This classification system is intended to bridge the gap between notions of equality, non-discrimination law, and decisions around how to measure fairness and bias machine learning. Bias transforming metrics are essential to achieve substantive equality in practice. I will conclude by introducing a bias preserving metric 'Conditional Demographic Disparity' which aims to reframe the debate around AI fairness, shifting it away from which is the right fairness metric to choose, and towards identifying ethically, legally, socially, or politically preferable conditioning variables according to the requirements of specific use cases.

Background Readings: "On the (im)possibility of fairness" Sorelle Friedler et al.

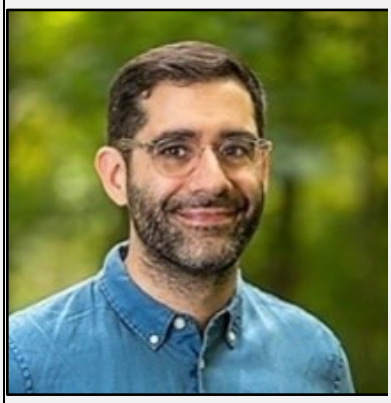
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

3.30.22



"Explanations in Whose Interest"

Salon Barocas

Principal Researcher, Microsoft Research (NYC lab)
Adjunct Assistant Professor, Department of Information Science, Cornell University
Faculty Associate, Berkman Klein Center for Internet & Society, Harvard University

Abstract: In the United States, the law requires that lenders explain their adverse decisions to consumers, one goal of which is to educate consumers about how to receive more favorable decisions in the future. Scholars have recently proposed a range of new techniques to help lenders realize this goal when their decision making relies on machine learning. However, attempts to directly map these techniques onto applications in finance are often somewhat stylized, failing to consider important aspects of lending in practice. As we will show in this paper, lending decisions are rarely binary (i.e., lend/don't lend). Machine learning models are often used by lenders to estimate consumers' risk of default, not to classify applicants as creditworthy or not; these estimates of risk inform a more complex decision about the terms on which lenders are willing to grant credit to consumers. Differences in the terms of a loan often result in very different utility for consumers and lenders. In fact, access to credit on unfavorable terms can be actively harmful to consumers, even if it might be profitable for lenders. Very little of the existing scholarship on explainable AI in finance—or that uses lending as a motivating example—takes these crucial details into account. As a result, many of the proposed methods for explaining adverse lending decisions may not help consumers achieve better outcomes—and may even harm them in some cases.

Background Readings: "The Use of ML for Credit Underwriting Market and Data Science Context" FineRegLab

Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

4.6.22

"Equality and Equity in Product Development"

Tulsee Doshi

Google Head of Product, Responsible AI & ML Fairness

Abstract: What does it mean to build a fair, inclusive, and/or equitable product? In this class, Tulsee Doshi, Head of Product for Responsible AI at Google will share how different product contexts merit different definitions and approaches to *fairness and* discuss how solving for *equality* may not in fact be solving for *equity.* As a class, we'll walk through some case studies and examples of different metrics & methods and highlight the importance of societal context in product development.

Background Readings: "Fairness and Abstraction in Sociotechnical Systems" Selbst, et al.
"Measuring Fairness" +PAIR Explorables

Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

4.13.22

"Ethics in the Metaverse"
Michael Brent

Responsible AI Expert, Boston Consulting Group

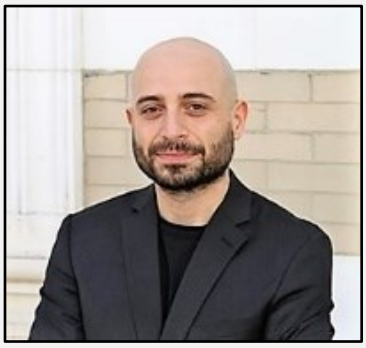
Cognitive Science Seminar: The Ethics of Artificial Intelligence

David Miguel Gray – Department of Philosophy
Dmgray2@memphis.edu

Guest Presentations 4:00 – 5:15 Central Standard Time (UTC -06:00)

https://memphis.zoom.us/j/7552850612?pwd=OFNUbVROS3Uxc1hjQ3E0L1BsTFZYQT09
Meeting ID: 755 285 0612
Passcode: 0ZUeBc

4.20.22



"Diversity in Sociotechnical Machine Learning Systems"

Sina Fazelpour

Assistant Professor of Philosophy & Computer Science, Northeastern University
Fellow, World Economic Forum's Global Future Council on Data Policy

Abstract: There has been a surge of recent interest in sociocultural diversity in machine learning (ML) research, with researchers (i) examining the benefits of diversity as an organizational solution for alleviating problems with algorithmic bias, and (ii) proposing measures and methods for implementing diversity as a design desideratum in the construction of predictive algorithms. Currently, however, there is a gap between discussions of measures and benefits of diversity in ML, on the one hand, and the broader research on the underlying concepts of diversity and the precise mechanisms of its functional benefits, on the other. This gap is problematic because diversity is not a monolithic concept. Rather, different concepts of diversity are based on distinct rationales that should inform how we measure diversity in a given context. Similarly, the lack of specificity about the precise mechanisms underpinning diversity's potential benefits can result in uninformative generalities, invalid experimental designs, and illicit interpretations of findings. In this work, we draw on research in philosophy, psychology, and social and organizational sciences to make three contributions: First, we introduce a taxonomy of different diversity concepts from the philosophy of science and explicate the distinct epistemic and political rationales underlying these concepts. Second, we provide an overview of mechanisms by which diversity can benet group performance. Third, we situate these taxonomies—of concepts and mechanisms—in the lifecycle of sociotechnical ML systems and make a case for their usefulness in fair and accountable ML.

Background Reading: "Algorithmic bias: Senses, sources, solutions" Sina Fazelpour & David Danks
"Discriminating Systems: Gender Race Power in AI" Sarah West et al.