Texas A&M University 5th Bovay Workshop on Engineering and Applied Ethics

Program

AI and the Ethical Engineering of Complex Systems

April 26, 2022

Location: YMCA – Department of Philosophy, Room 401 **Schedule:**

8:30-9:00 Breakfast (provided)

9:00: **Opening remarks**: Dr. David Koepsell, Texas A&M University

9:15: **Rune Nyrup**, Cambridge University

"Towards an Ethics of Explainable Artificial Intelligence"

A common objection to the use of machine learning (ML) in high-stakes decision making is that ML models risk becoming in some sense 'uninterpretable' or 'unexplainable'. Explainable AI (XAI) is a subfield of AI research that seeks to develop technical tools for overcoming this challenge. These include tools for extracting and displaying certain kinds of information about ML models or their outputs (e.g., simplified model representations, representative training data, most salient input variables, uncertainty estimates), or techniques for ensuring that ML models are trained in a way that renders from intrinsically easier to understand. However, contrary to dominant assumptions in this literature, explainability is not an unqualified ethical good or even an ethically neutral tool. Powerful methods for generating convincing-sounding explanations or seemingly intuitive models could be used for pernicious purposes, e.g., to manipulate or deceive. In this paper I make three contributions towards addressing this concern. First, I outline a general framework for analysing explainability claims, based on broadly pragmatist theories from recent philosophy of explanation and understanding. Second, I use this framework to argue that even when no actor has malicious purposes, XAI techniques could inadvertently generate misleading explanations, especially in contexts where the values and interests of explainer and audience diverge. Finally, I outline some proposals for future research directions to help us better evaluate the ethics of XAI tools

10:15: **Jobst Landgrebe**, State University of New York (SUNY), University at Buffalo

"The political philosophy of AI"

"Power is a fundamental relationship in all human communities and societies. Political philosophy arose in Greece as power structures became institutionalized and social norms were turned into law. Will Artificial Intelligence change the way in which rulers can exercise power over their subjects? I review the leading theories of political power from Hobbes to Popitz before using the framework of Poptiz' 'Phenomena of Power' to analyze how AI will affect the four fundamental types of power - namely power of action, instrumental power, authoritative power and data-setting power. Finally, I apply the results of this analysis to the emerging global governance, a new type of global centralized form of power."

11:15-11:30: **Coffee Break** (Provided)

11:30: **Barry Smith**, University at Buffalo

"The Machine Will"

Deepmind's AlphaFold is a great leap forward in the attempts by biologists to address the 50-year-old challenge of predicting a protein's structure from its amino acid sequence. In its inventiveness and complexity, and in the degree to which it draws on so many disciplines and on so much prior work on the part of earlier generations of human beings, Alphafold is comparable to some of the most remarkable achievements of humankind. It could not, of course, have been created without the computer and the advances in machine learning which the computer made possible. But as the documentation of how it was built makes clear, ***AlphaFold is a product of the human will*.** My talk will address the meaning of this statement.

12:30-2:00: **Lunch** (provided)

2:00: **Peter Zuk** Center for Bioethics, Harvard Medical School

"Neural Data: Not for Sale"

Adrian Walsh argues that the commercialization of a good tends to corrode recognition of its intrinsic value or that of related entities. I consider how this claim applies to the collection and use of neural data. I argue that non-medical commercial uses of such data do not merely tend to corrode recognition of the intrinsic value of persons, but essentially involve a failure of such recognition.

I begin by cataloguing some views of the relation between a person and neural data derived from her brain, ranging from the extreme *identity view* apparently endorsed

by some transhumanists, to the more plausible *relationist* and *separatist views*. On all of these views, it turns out that collection and use of neural data involves taking, toward the individual from whom data is derived, what P.F. Strawson calls the objective attitude. That is, it involves treating them as a thing that can be described and intervened on mechanistically. Interventions on the brain are in this respect a special case; we needn't take the objective attitude toward the whole person to intervene on a broken leg or ruptured spleen. But we do when intervening on the brain, given plausible accounts of the brain-mind relation and the mind-person relation.

Neural data has both medical and non-medical applications. Neural data derived from invasive recordings is crucial to the development and refinement of neuromodulation therapies such as deep brain stimulation. The objective attitude is permissible (and desirable) in medicine and medical research because medical interventions aim at restoration of a person's health, well-being, and/or autonomy. Thus, despite regarding the individual as a site of mechanistic intervention, the intervention also simultaneously regards them as an end *qua* person.

However, taking the objective attitude toward a person for purposes of selling a product is incompatible with valuing them as an end. Here we can distinguish two broad nonmedical uses of neural data: use in the *selling* itself (so-called "neuromarketing") and use in the *product* sold (as in consumer neuromodulation devices). In both cases, the objective attitude is taken in the absence of the concern for the person as an end that, in medicine and medical research, render taking it permissible (and desirable). The person is viewed *merely* as a mechanistic system for intervention (and resulting profit). This, plausibly, is the kernel of truth in otherwise ambiguous concerns that novel neurotechnologies will result in dehumanization. But here we see that it is not the technologies as such that are objectionable, but instead their deployment for aims that make no essential reference to the good of those who are intervened upon.

I close by addressing the objection that these claims presuppose a controversial Kantian moral theory. To the contrary, proponents of other moral theories (as I myself am) can endorse these claims as long as their theories allow for goods and bads (or rights and rights-violations, or virtues and vices, etc.) that constitutively involve relations of a certain kind. And there are independent reasons for thinking that one's moral theory should do this.

3:00: **Martin Peterson**, Texas A&M University

"AI and Value Alignment: A Quantitative Measure"

Abstract: It has been suggested that AI systems should be designed to align with the moral values of the AI user. In this talk I address the following question: How can we *measure* the degree to which an AI system aligns with a set of moral values specified by a user? My point of departure is Peter Gardenfors' theory of Conceptual Spaces, in which concepts are defined as convex regions in a multidimensional similarity space defined by one or several prototypes. In previous work I have argued that moral principles can be construed as conceptual spaces in the manner proposed by

Gardenfors, and that a significant advantage of this "geometric" approach to applied ethics is that it enables ethicists to sharpen discussions of moral principles in ways that have previously been beyond the limits of the discipline. The point of departure for my measure is Aristotle's formal principle of justice, according to which we should "treat like cases alike." (NE 1131a10-b15.) I suggest that it is plausible to maintain that the more similar a pair of moral choice situations are, the more reason do we have to treat them alike, which leads to the following observation: (i) if two cases x and y are fully similar in all morally relevant aspects, and if principle p is applicable to x, then p is applicable to y; and (ii) if some case x is more similar to prototype y than to any other prototype z, and p is applicable to y, then p is applicable to x as well. This suggests a linear measure of value alignment that counts not just how often, but also the extent, to which a moral principle accepted by the user is violated by comparing how similar the output of the AI is to the similarity judgements reported by the user.

3:50 - 4:15: Coffee Break

4:15 **Austen McDougal**, Stanford University

"Cheap Versus Significant Considerateness for Extended Minds"

There is a positive form of attention that I call considerateness: taking cherished facts about others into consideration in daily life, even when that has little instrumental value. Reminders on social media that facilitate considerateness include birthday reminders, email reminders (e.g., Gmail reminds you that you have not responded), photo reminders (e.g., Instagram reminds you of a happy memory with friends on this day two years ago), and others. While reminders facilitate an awareness similar to what one might have in a caring relationship, they undermine the significance of considerateness, which is precisely that one cares about someone enough to take them into mind. I call instances of considerateness *cheap* when they don't require the relevant sort of caring attitudes—in the above cases, due to the effortless way in which our minds are extended by current social. In the second half of the talk, I consider a range of remedies that might be taken. Rather than a default opt-out setting, information and reminders on social media could be opt-in, to varying degrees of deliberateness. Separately, social media could also facilitate a range of expressions of care, which make up for the lack of significant considerateness.

5:05: Berit "**Brit" Brogaard**, University of Miami

"Human Brain Organoids: Scientific and Ethical Implications"

Human brain organoids are three-dimensional, self-organizing neural structures grown in vitro from human pluripotent stem cells. The advent of lab-grown miniature brains have ignited a great deal of enthusiasm in the neuroscientific community, as they have the potential to accelerate research on human brain development, function, and

disease. However, the increasing maturity and complexity of human brain organoids raise a host of ethical and regulatory issues pertaining to their moral status, engineering, and storage, transplantation into non-human animals, and potential integration with deep learning artificial neural networks. In this talk, I begin with a brief overview of the status quo of brain organoid research and the potential of these brain structures for rudimentary consciousness. I then review the most realistic future applications of brain organoids in AI and robotics. Lastly, I address ethical issues pertaining to the use of potential human brain organoid-machine chimeras for the purpose of human servitude.

6:30: Dinner at Solt, 830 University Dr E #400, College Station, TX 77840