Machine Wisdom Workshop 2

Held at the University of Pittsburgh, Cathedral of Learning, May 12-14, 2022

With support from…

Templeton World Charity Foundation

Pittsburgh Center for Philosophy of Science

Abstracts booklet

Thursday May 12

	Names	Titles	Session chair
1:00-1:15	Brett Karlan & Colin Allen	Welcome Address	
1:15-2:00	John Sullins, Sonoma State University	Machine Wisdom, Artificial Phronesis, and Inner Speech	Colin Allen
2:00-2:45	Beba Cibralic, Georgetown University	Responsibility for Speaking Machines	Conny Knieling
2:45-3:30	***BREAK***		
3:30-4:15	Igor Grossmann, University of Waterloo	Psychological Science of Wisdom	Eleni Angelou
4:15-5:00	Matt Stichter, Washington State University	Flourishing Goals, Metacognitive Skills, and the Virtue of Wisdom	Jamie Kelly
5:00-5:45	A.G. Holdier, University of Arkansas	Consider the Lobster: Speciesism, Algorithmic Bias, and the Ethics of Care	Serife Tekin
7:00-9:00	Dinner for all speakers at the Wyndham Pittsburgh University Center		

Friday May 13

9:00-9:45	Jamie Kelly, Vassar College	Marx, Robots, and Social Reproduction	Micah Musser
9:45-10:30	Michael Barnes, Rotman Institute of Philosophy, Western University	Working for the Machine	Dasha Pruss
10:30-11:00	***BREAK***		
11:00-11:45	Shannon Vallor, University of Edinburgh	Moral Mirrors and Telescopes: The Opportunity for AI-Augmented Wisdom	John Sullins
11:45-12:30	Brian Tebbitt, University	Value Frames and the Godlike	Konrad Werner

	of Minnesota	Position	
12:30-2:00	***LUNCH (on your own)***		
2:00-2:45	Ravit Dotan, University of Pittsburgh	g	Heather Douglas
2:45-3:30	Micah Musser, Center for Security and Emerging Technology, Georgetown University	How to Regulate AI Models	Michael Barnes
3:30-4:00	***BREAK***		
4:00-4:45	Kathleen Creel, Stanford University	Picking on the Same Person: Artificial Judgment and its Discontents	Xin Hui Yong
4:45-5:30	Michael Tamir, University of California, Berkeley & Elay Shech, Auburn University	Understanding and Deep Learning Representation	Brendan Fleig-Goldstein
5:30-6:15	Chris Davison, Ball State University	The Ethical, Multimodal, Modeling, and Adaptive (EMMA) System	Brett Karlan

Saturday May 14

9:00-9:45	Sina Fazelpour, Northeastern University	Disciplining Deliberation	A.G. Holdier
9:45-11:15	"Pittsburgh Blitz Session" • Xin Hui Yong: A Seat At The Table? Modelling Whether Algorithmic Agents Compound the Dampening of Minoritized Voices • Conny Knieling: Arbitrariness in Algorithmic Decision-Making as a Moral Problem • Konrad Werner: The Machine Wisdom of Not Being Too Wise: Social Apps and Cognitive Confinement in the Time of Mental Health Crisis		Brett Karlan
11:15-11:30	***BREAK***		

11:30-12:30	Brett Karlan & Colin Allen	Future Directions and General Discussion

Michael Barnes (Western University, Philosophy)

Working for the Machine

There is a long history of some people (usually the more powerful) taking credit for the work of some other people (usually the less powerful). Pharaohs, Kings, or other Heads of State claim responsibility for building pyramids, monuments, and railways, when the real labor was, for the most part, performed by slaves, serfs, and other impoverished laborers. In general, the managers claim credit for the long hours the rank-and-file workers puts in. And sometimes this involves a bit of deception. Thomas Jefferson, for example, installed a series of dumbwaiters in his Monticello home and marveled his guests as food and wine—prepared in another room by slaves—appeared when needed, as if by magic.

A more recent development of this old pattern is the work of human laborers being passed off as that of 'artificial intelligence.' This comes in many forms, including so-called 'autonomous delivery drones' that are, in actuality, piloted by workers in office parks across the world. This example is representative of a larger fact: AI is changing the working conditions of human employees—or, more commonly now, independent contractors—and not simply by replacing them, but often by asking them to hide themselves and their work. This type of obfuscation (or straightforward lying) is usually done either to make the technology seem more impressive to outsiders, or, just as often, to hide or render invisible the working conditions of the laborers for one reason or another.

Mary Gray and Siddharth Suri coined the term 'ghost work' to capture these types of working conditions, and it is a growing phenomenon in our increasingly (at least seemingly) automated and (most assuredly) globalized world. And as Phil Jones, author of Work Without the Worker, has reported, some of this ‘click work’ is even being pushed as a ‘solution’ to the problem of global poverty, and is sometimes targeted at people living in refugee camps.

The proliferation of ‘ghost work’ / ‘click work’ is therefore the hidden underbelly of the contemporary explosion of AI systems, which is simply to say that these systems rely on countless human workers to do the job of labeling, training, correcting, and sometimes even imitating ‘Artificial Intelligence.’ And, most importantly, this is happening with very little attention, oversight, or even awareness.

This paper aims to investigate the low-wage work that underpins so much current technology and makes a first pass at proposing the theoretical foundation for some useful guidelines for effective ethical oversight. Taking inspiration from the anti-sweatshop movement of the 90s, I consider how ‘transparency’ may be understood in this context. Rather than locating the factories and opening up the doors to independent inspectors, transparency here seems to require something else. It may require, for example, workers knowing what they’re working on— something that is currently often impossible. And beyond the workers themselves, outside groups should be permitted a better understanding of where the data that powers AI systems

are coming from and how they are being put to use. Figuring out what actual oversight would look like here, and who are the correct people who ought to be overseeing these things, is a useful and urgent project.

I approach this topic by examining the situation of content moderation, which is another instance of a (relatively) invisible labor force supporting a modern technological wonder: global social network platforms.. Currently, thousands of human content moderators are employed by the many technology companies who profit off user-generated content posted online. Big Tech relies on these moderators—often employed via third-party mediators—to sift through countless toxic posts every day. Most of these moderators are underpaid contractors, often overseas, suffering from psychological harms behind Non-Disclosure Agreements. Yet they are essential workers in the global information supply chain—and the harms they experience are some a full account of online life must explore.

For both click work in general, and content moderation in particular, the question of 'Artificial Intelligence's capacity to both improve the working conditions of human laborers is severely put to the test. And content moderators demonstrate the tensions within this dynamic well. This is because Mark Zuckerberg (CEO of Meta, formerly known as Facebook) has repeatedly claimed that 'AI moderation' will be crucial in the (near?) future to keep the platform safe—though he seldom ever mentions the moderators who do that work currently. And, of course, these future AI systems that Zuckerberg dream of will be built upon the work of those human moderators whose byproduct is the data used to train AI. As such, they've been tasked with the potential goal of replacing themselves.

AI therefore poses a problem even for those who, for various reasons, may wish to defend the general practice I started with. That is, some might the key work is done by those who design the plans, and the laborers merely carried their wishes out. In the case of AI, however, the 'brains' are the result of, among other things, the efforts of many low-level laborers who cleaned and labeled the data that its algorithms could later process. So, in addition to potential labor exploitation, there seems to be a case for intellectual exploitation as well. After all, the reason AI systems aren't currently suitable to perform these tasks is that they require a human touch.

In sum, in this paper, I consider the labor exploitation faced by the invisible workforce that powers AI alongside the propagandistic way in which AI is discussed by the leaders of Big Tech. As we face the specter of automation—or fauxtomation—questions over the purpose and value of work are more important than ever, and it is necessary that we face these with a clear understanding of how this technology operates in a globalized world. And beyond clearing away the various obfuscations, we must seriously consider how to improve the conditions of a growing workforce, if, for no other reasons than the fact that we may be joining them soon.

Beba Cibralic (Georgetown University, Philosophy)

Responsibility for Speaking Machines

The purpose of this talk is to explore the ways in which machine learning systems like GPT-3 will challenge how we establish responsibility for speech and expression. More specifically, I will answer the question, whom do we hold epistemically responsible for the speech of machine agents? In order to answer this question, two intermediate questions need to be addressed. First, do certain machine agents count as speakers? Second, how does epistemic responsibility differ from moral responsibility? Having answered these questions, I explore what it would mean to count GPT-3 as a speaker and hold it epistemically responsible for its output. Ultimately, I argue that if machine agents do count as speakers who can be held epistemically responsible, it is even more important to regulate the creation of these machines.

Kathleen Creel (Stanford University, Philosophy)

Picking on the Same Person: Artificial Judgment and its Discontents

Good judgement involves the proper application of categories and judgement of people. Human mistakes are inevitable, but fortunately heterogenous. Not so with machine judgement. Using the same machine learning model for high-stakes decisions in many settings amplifies the strengths, weaknesses, biases, and idiosyncrasies of the original model. When the same person re-encounters the same model again and again, or models trained on the same dataset, she might be wrongly rejected again and again. Thus algorithmic monoculture could lead to consistent ill-treatment of individual people by homogenizing the decision outcomes they experience. This talk will formalize the measure of outcome homogenization, describe experiments on US census data that demonstrate that the sharing of training data consistently homogenizes outcomes, then present an ethical argument for why and in what circumstances outcome homogenization is wrong.

Ravit Dotan (University of Pittsburgh, HPS)

Participatory AI Governance

The last few years have seen more and more AI ethics scandals. Time after time, we've seen that AI algorithms have exacerbated bias and discrimination, have been associated with breaches of privacy, and have caused other harms to people and the environment.

What can we do to mitigate the adverse effects of AI systems and harness their power to create positive impacts on people and the environment? Currently, the prominent approach focuses on generating AI ethics principles, which regulate the design and functionality of AI systems with an eye towards mitigating their potential harms. Here are a few typical examples, to give a taste of what they look like: "AI must be designed to protect user data and preserve the user's power over access and uses" (IBM), "The Department will take deliberate steps to minimize unintended bias in AI capabilities" (US Department of Defense).

Articulating AI principles is increasingly popular, as evidenced by hundreds of sets of AI ethics principles, written by governments, inter-governmental organizations, corporations, non-profits, and academics, typically focusing on topics such as transparency, fairness, and privacy.

Why do we formulate AI ethics principles? One hope seems to be that they would contribute to developing AI systems that positively impact humans and the planet. For example, Australia's Artificial Intelligence Ethics Framework states that "Australia's 8 Artificial Intelligence (AI) Ethics Principles are designed to ensure AI is safe, secure and reliable". The principles are said to help "achieve safer, more reliable and fairer outcomes for all Australians", "reduce the risk of negative impact on those affected by AI applications" and "businesses and governments to practice the highest ethical standards when designing, developing and implementing AI."

A minority of organizations also develop tools for implementing these principles, e.g., checklists, questionnaires, computational tools to reduce bias in datasets, and workshop-style events for raising awareness in design teams. The tools are used in all stages of the AI lifecycle and they can be helpful in moving the field of AI ethics from talk to action.

Yet, most AI ethics principles continue to remain abstract and suffer from problems of efficacy in their implementation, even in the presence of tools and techniques that help to operationalize those principles from a technical perspective.

We argue that effective AI ethics governance, that accomplishes goals like the ones the Australian Government aspires to, requires two additional components, on top of traditional AI ethics tools and principles.

First, effective AI ethics governance requires external audits of the organization's ethical engagement. The audit we propose includes two fundamental questions, which we break down to six organizational aspects that auditors should focus on:

How well integrated are AI ethics tools and principles into the organization's workflow?

> **AI ethics Everyday Practices** - Are AI ethics tools and principles effectively integrated into the everyday practices of the employees across the AI development pipeline?
>
> **AI ethics Incentives** - Are the organization's financial and performance incentives across the AI development pipeline aligned with the organization's ethical aims?

Do the organizational structures and procedures support raising and responding to ethical critique?

> **Criticism Uptake** - Is there uptake of ethical criticism with regard to the AI systems it develops and uses?
>
> **Critique Solicitation** - Does the organization cultivate and seek ethical criticism about the AI systems it develops and uses?
>
> **Literacy Cultivation** - Does the organization cultivate the literacy required to engage with AI ethics?
>
> **Access to Diverse Critique** - Does the organization have access to diverse perspectives from all stakeholders (internal and external) in its AI ethics discussions?

Second, we argue that effective AI ethics governance also requires that the results of these audits be accessible and publicly available. We also argue that achieving these goals requires that the results of the audit are machine-readable, archivable, version-controlled, and standardized.

We make the case for how public audits of organizational ethical engagement can not only increase the likelihood that AI ethics principles and tools will achieve their stated goals, but also help rectify power imbalances.

Chris Davison (Ball State University, Computer Technology)

The Ethical, Multimodal, Modeling, and Adaptive (EMMA) System

The objective of this research and testbed is to engineer an Intelligent Physical System (IPS) in the form of a building energy management system (BEMS) imbued with AI ethical reasoning: the Ethical Multimodal, Modeling, and Adaptive (EMMA) system. Most current efforts to integrate AI into BEMS focus on achieving cost savings within standardized temperature and humidity ranges. These efforts ignore fairness, bias, and inclusivity issues that arise in the context of a BEMS, such as those arising from individual differences and the various roles different tenants play in buildings—roles which tend to correlate with gender and other social and demographic characteristics. Failure to take such differences into account can disproportionately bias the system against certain individuals and groups.

This joint project between Ball State University (BSU) and University of Pittsburgh (Pitt) is funded by Templeton World Charity Foundation (TWCF) and investigates using tenant-supplied comfort profiles and other behavioral data to enhance ethical decision making within a joint sociotechnical system.

Enhancements to our previously-prototyped EMMA system takes information provided by users of the EMMA mobile App and combines it with multi-modal sensor streams taken in real time from a BEMS on the BSU campus as well as usage and cost modeling information from utilities data. Tenant preferences, resource conservation considerations, and the HVAC system's own survival and maintenance imperatives factor into EMMA's responses to tenants' comfort requests. These factors also play into EMMA's eXplainable Artificial Intelligence (XAI) methods and into evaluating EMMA's capacity to meet Fairness in AI (FAI) objectives.

This research testbed represents the first known attempt to integrate human ethics and machine ethics into a BEMS to create a socio-technical, joint decision-making system. We aim to create the first functional artificial moral agent (AMA) capable of balancing societal good alongside human needs and desires. Moreover, we are pursuing a novel approach to FAI: fairness monitoring by a machine surrogate that is capable of preserving and independently analyzing more details than non-expert users can process. There are several other innovative contributions: We investigate and analyze FAI and XAI strategies, such as the effects of EMMA's explanations, with different types of ethical framing presented at different levels of detail, upon tenants' subsequent explanation requests and interactions with the EMMA App. Our working hypothesis is that fair outcomes are enhanced when EMMA presents explanations about other users' needs and preferences. Finally, our project will create an open dataset of user interactions that we will use to build better agent based models (ABMs), thus creating an experimental testbed for BEMS interventions affecting FAI outcomes.

For this workshop, we would like to explore potential problems with our IPS. Where could potential problems or pitfalls manifest? How can we anticipate and mitigate these problems? An example of this is unfair or unethical outcomes. Is it possible to anticipate ethically questionable outcomes?

A potential source of wisdom by design is the use of focus groups at all stages of the development lifecycle. Is this a plausible approach? Which constituent groups are invited and what do these focus groups look like and how do they perform?

Trust is another major concern in AI. Our approach was to create a personal surrogate, adopting the fight fire with fire approach. How and why should people trust their own surrogate to identify bias? Will the surrogate actually intercede on behalf of the user given it is a part of the whole EMMA system? How and why should humans trust AI? How do we engineer trust into a system? Our approach was to involve ethicists and philosophers from the beginning to drive the development of the system. Is this a reasonable approach? How can we improve? Also, we incorporate a sandbox idea (benign environment) where limits testing and what-if scenarios can be modeled. Is it possible to identify outlier conditions from the sandbox?

Finally, the tool we used for modeling was eQuest. This tool is an evolution of the DOE-2 tool for energy consumption. eQuest modeling is accepted for building LEED certification (most widely adopted green energy rating system). We found several issues with eQuest in the process of designing EMMA. First, it was not granular down to smaller time frames below one hour. This impacts the granularity of the AI feedback to a close approximate cost of the preference request, but not an exact cost. This less than exact feedback to the user may well influence the user's decision.

The second issue with eQuest was the lack of heatmaps. A change in one office will impact adjacent offices over time. This is not captured by the modeling software. This has ethical implications if adjacent office tenants have medical issues or other conditions impacted by temperature and humidity.

Further questions for the workshop:

- How can the development of practical expertise(through limits testing and other experience with the system) improve XAI and FAI outcomes in AI?

- What are the outcomes we want to improve and how do we measure them in the first place?

- Does focusing on wisdom and expertise unnecessarily ignore the important social problems and injustices intimately bound up with the rise of AI?

- How can we encode expertise and wisdom in these systems during all development stages?

Sina Fazelpour (Northeastern University, Philosophy and Computer Science)

Disciplining deliberation: The case for exercising caution in interpreting formal trade-offs

Formal analyses of decision scenarios routinely uncover trade-offs between different desiderata of value to decision-makers, and offer a precise perspective for deliberating about these value conflicts. This paper focuses on two such highly publicized trade-offs in machine learning (ML)—accuracy-fairness and accuracy-interpretability. In many cases, these abstract trade-offs are taken to be directly applicable to practical settings by researchers, practitioners, and policy-makers alike, forming a core focus of normative engagement with the relevant domains. I argue against this attitude, and offer four sets of reasons for exercising caution in interpreting the practical applicability of these trade-offs, pertaining to their (i) semantic, (ii) empirical, (iii) compositional, and (iv) dynamic significance. I show how neglecting these can restrict our normative engagements with and aspirations for predictive technologies. Together, these considerations provide a diagnostic lens for reasoning about the epistemic and ethical significance of these trade-offs in the evaluation, design, and governance of ML-based decision-making.

Igor Grossmann (University of Waterloo, Psychology)

Psychological Science of Wisdom

In a time of disagreements about values, politics, and cultural practices, psychological scientists have turned to possible antidotes to societal acrimony – the concept of wisdom. Diverse approaches to defining wisdom in psychology exist. I will describe what social scientists studying wisdom see as common across a myriad of definitions: epistemic humility, consideration of multiple perspectives and ways a situation may unfold, observer viewpoint on a situation, and willingness to be open-minded to different perspectives. Not all measurement approaches are equally suitable to capturing this meta-cognitive architecture of psychological wisdom. At the end, I will outline potential benefits of considering these characteristics in the context of interaction with and creation of artificial intelligence. Instead of a reductionist focus on abstract rationality, psychological wisdom research suggests that adaptive artificial systems ought to focus on multiple unknowns and the priority of the particular.

A.G. Holdier (University of Arkansas, Philosophy)

Consider the Lobster: Speciesism, Algorithmic Bias, and the Ethics of Care

Discussions of the ethics of algorithms, AI, and related technologies generally focus on questions relevant to actual or potential users, such as how they are (or could be) harmed, wronged, or otherwise influenced. Though important, this emphasis on considering human relationships with AI threatens to ignore the many additional nonhuman subjects affected by modern technology; in this paper, I contend that the rights and interests of nonhuman animals should be given more weight in contemporary analyses of emerging technologies and demonstrate how doing so can help to foster a deeper appreciation for the motivations of AI ethics in general.

Most explicitly, nonhuman animals can be directly affected by AI technologies (intentionally or otherwise) in ways that threaten their safety, such as in baseline experiments for neural net development (Bossert and Hagendorff 2021, pg. 2), programming decisions for unmanned aircraft systems and self-driving cars (Bendel 2016, pg. 107), or via potential existential risk in long-term calculation scenarios (Ziesche 2021). But this paper focuses on a heretofore-unexplored problem for nonhuman animals confronted by uncaringly speciesistic human cultures: the amplification of speciesism via algorithmic bias.

Algorithmic bias is a "systematic deviation in algorithm output, performance, or impact, relative to some norm or standard" (Fazelpour and Danks 2021, pg. 2); when the relevant standards are moral norms, as when algorithms encode and recapitulate systemic social injustices (Danks and London 2017, pg. 2), then algorithmic bias poses ethical dangers. Johnson (2021) has identified how algorithmic reliance on proxy attributes — "seemingly innocuous attributes that correlate with socially-sensitive attributes" (pg. 9942) — contributes to the emergence of discriminatory patterns in AI outputs; calling this the Proxy problem, Johnson 2 points out that algorithmic effectiveness in oppressive social environments seemingly requires those algorithms to recapitulate that oppression (pg. 9957): the double-bound algorithm is either subject to immoral bias or it fails to render effective decision-making procedures within the immoral environment.

I contend that the dataspace of a speciesistic culture like ours is suffused with measurements of proxy attributes that ultimately help maintain unjust ideologies harmful to nonhuman animals. For example, consider how internet searches for "restaurant locations," "party supplies," "exercise plans," or "holiday traditions" might all (for various algorithmic reasons) output results prioritizing carnistic habits — despite none of these queries requiring meat-consumption a priori. Although search engines do not directly rely on the commercial animal feeding operations (or "factory farms") that torture and kill billions of nonhuman animals each year, the speciesistic bias embedded within their algorithms serves to buttress the immoral practices of factory farms by further normalizing the products of their operation. Furthermore, the algorithmic interpellation of carnism hinders the possibility of effective policy problem construction and implementation by defenders of animal welfare interested in regulating

factory farms; these sorts of unintended social externalities are a key factor that contemporary philosophers of technology must consider more carefully.

Finally, in addition to considering how implicit technological biases disadvantage nonhuman animals, this paper also considers how AI technologies can, by explicitly promoting uncaring responses to our fellow creatures, inhibit the robust cultivation of what Vallor (2016) calls technomoral virtues. Vallor argues that character traits like empathy (or compassion), misericordia (or pity), and care are not only important facets of human experience (particularly in social situations), but ones that take on new dimensions in the technosocial environment we inhabit; as she explains, "an enriched understanding of 21st century care must intelligently incorporate the assistance of technology rather than dismiss it out of hand" (pg. 140).

But consider technology like the facial recognition software designed to detect "stress" in factory-farmed animals and alert attendants quickly about creatures in need of attention (Bridgeman 2021); while such technology might be defended as a harm-reducing measure, this kind of "humane-washing" of the fundamentally immoral environment of the factory farm fails to promote genuine nonhuman animal flourishing (as if any inhabitant of a CAFO might not feel "stress" at their predicament?). Furthermore, by replacing authentic creaturely interaction with a technological intermediate, this facial recognition software removes an opportunity for the human attendants to connect with and thereby genuinely care for the nonhuman animals under their watch. In this way, AI technology not only further subtends the unjust, speciesistic objectification of nonhuman animals as mere consumable resources, but it also violates importantly humane responsibilities to care for fellow creatures in tangible ways.

Nearly twenty years ago, David Foster Wallace asked the readers of Gourmet Magazine to "Consider the Lobster" by waxing eloquently about the ethical ramifications of the Maine Lobster Festival; confronting his audience with questions about the morality of their gustatory choices, Wallace cheekily wondered whether people's "refusal to think about any of this [is] the product of actual thought, or is it just that they don't want to think about it?" (2005, pg. 166). The argument of this paper is that speciesistic algorithmic bias now means that we have even more ways to avoid caring about such questions at all.

Jamie Kelly (Vassar College, Philosophy)

Marx, Robots, and Social Reproduction

In this paper I develop a roughly Marxist account of artificial labour. That is, I modify Karl Marx's account of labour from Capital, Volume One, in order to show how an Artificial Intelligence could become a bona fide labourer (rather than a mere machine that serves to increase the productivity of human labour). I then go on to examine the potential consequences of artificial labour for Marx's account of economic production. I conclude that in order for artificial labour to function in a way that is analogous to human labour, robots would need to engage in social reproduction: they would need to use some of their labour to maintain or reproduce themselves into the future.

In the first part of the paper, I give a summary of Marx's account of labour aimed at a general audience, which presupposes no prior experience reading Capital. This results in a provisional definition of labour as "purposeful activity aimed at the satisfaction of human need." (from Marx, Capital, chapter 7). I then use that general definition of labour to argue for the possibility of artificial labour: robots capable of purposive activity aimed at satisfying human needs. After establishing the possibility of artificial labour within Marx's account, I lay out a number of vexing questions about how artificial labour would fit into existing social and economic relationships, and in particular regarding whether artificial labour could serve as a source of profit within capitalism. I examine four possibilities:

1. The Simple Story
On this account, artificial labour would behave just like a new supply of cheap human labour. If purposive activity is all that matters for economic and social relations, then robot and human labour should be interchangeable. I provide a number of reasons for rejecting this account.

2. The Incommensurability View
According to this view, artificial labour is incommensurable with human labour, and so robots cannot be a source of profit. This view denies the possibility that robots could ever do anything more than increase the productivity of human labour. I provide a number of reasons for rejecting this account as well.

3. The Analogy with Colonial Slavery
In order to shed light on these issues, I clarify Marx's stance on how slave labour interacted with wage labour during the colonial era, and use that interaction as a template for thinking about artificial labour. On the account I develop here, artificial labour cannot be the source of capitalist profits, but it can produce commodities that feed into capitalist production. I argue that this account provides some helpful insight into how robots might fit into production, but that it cannot answer some fundamental questions about artificial labour.

4. The Social Reproduction Account

Finally, I argue that in order for artificial labour to create profit within capitalism, robots must be capable of reproducing themselves. That is, in order for a robot to function like a human labourer within existing social and economic relations, robots must use some of their labour to maintain or reproduce themselves. I argue that this account provides a plausible explanation of how artificial labour might function, and tease out some important insights for our thinking about robots in general.

In order to motivate the Social Reproduction account, I work through several thought experiments intended to demonstrate how artificial intelligence could come, over time, to more closely resemble human labour. One of my key fictional examples is Hopper:

> After their tech startup goes bankrupt, software developers release Hopper, their new AI, onto the internet. Hopper survives by hiding out in various cloud services, borrowing idle computing power, and building a shadow network for itself. In order to pay for its remaining hardware needs, Hopper pretends to be human and works "remotely" as a data analyst.

I argue that Hopper provides an important threshold case: capable of artificial labour, and forced to engage in social reproduction, Hopper initially appears to be a plausible candidate for being a source of capitalist profit. I argue, however, that the deception built into this example, coupled with Hopper's lack of legal ownership rights, means that it cannot really replace human labor within our economy. I conclude that the creation of AI labourers would require a great deal more than just robots capable of purposive activity: it requires both self-reproduction, and legal ownership rights. Marx's account thus enables us to tease apart some of the key technological challenges facing artificial intelligence from the social and legal changes that would also be required to replace human labour with artificial labour.

Micah Musser (Center for Security and Emerging Technologies, Georgetown University)

How to Regulate AI Models: Contrasting a "Best Practice" Model of AI Regulation with a Focus on Impacts

Under most dominant learning paradigms, AI models learn appropriate weights and biases for some (usually narrowly scoped) task by optimizing for a single metric given by a loss function. Unfortunately, while extremely effective at encoding information about real-world domains, this approach makes it difficult to incorporate non-accuracy-related measures of "acceptable" performances into the model training process. Frameworks such as the "Fairness, Accountability, and Transparency" schema attempt to articulate these separate values, with conferences such as the ACM FAccT exploring how to embed these values in machine learning models themselves.

My talk aims to critique some of the existing approaches to embedding these types of values into machine learning systems, and to offer an alternative approach to that dominant in the machine learning literature that could provide policy recommendations with more traction. I proceed in three phases: First, I introduce a distinction between practice-based regulations of AI algorithms and impact-based regulations. Second, I argue that in some salient cases, although most of the technical literature focuses on identifying practices that can embed various social values into AI systems, this approach is mistaken, and that impact-based rules could both offer greater regulatory traction while often better satisfying the value in question. Finally, I attempt to articulate a general schema that can be used to identify what domains and contexts are suitable for this type of impact-based regulation.

I suggest that most machine learning research on AI and social impact focuses on the attempt to find general-purpose practices that can result in models that are, in themselves, more "fair," "transparent," "auditable," and so forth. Broadly, the types of practices that are the focus of this research can be divided into three classes: data manipulation, algorithmic adjustment, and the building of additional tooling. Data manipulation consists either of altering the composition of a data set (e.g. by oversampling from underrepresented populations) or of feature engineering (e.g. by dropping features that a model could associate with protected classes). Algorithmic adjustment can take many forms but generally operates with the goal of identifying a loss function or training mechanism that can cause a model to prioritize certain socially valuable attributes; for instance, using reinforcement learning to "penalize" a pre-trained language model for outputting sexist statements. Additional tooling is often related to the goal of transparency, e.g. by constructing tools that can be used to inspect the features that cause specific neurons to activate in deep neural networks.

Often, people who focus much of their technical research on developing these types of tools are inclined to think that AI regulation should focus on mandating compliance with certain practices, e.g. by mandating representativeness in training data. By analogy, this model of

regulation can be compared to food safety regulations: restaurants are not regulatorily obligated to only cause fewer than a certain maximum number of cases of food poisoning; rather, they are obligated to abide by certain best practices that demonstrably reduce the risk of customers being poisoned.

However, this technical focus on identifying "best practices" can sometimes be misguided. In this talk, I want to suggest that the technical efforts associated with data manipulation, algorithmic adjustment, and building additional tooling are, in many cases, not necessary to permit effective regulation of machine learning models in line with values such as fairness and accountability. In fact, I argue that these efforts are often conceptually confused in such a way that makes the stated goals of research impossible to deliver, while simultaneously creating a moral hazard problem that allows policymakers to wait for impossible technical improvements before acting.

My thesis can be summed up as follows: if the outcomes that would be consistent with a given value (e.g. lack of bias) within a given domain (e.g. criminal sentencing decisions) can be formally articulated, then compliance with the value can often be ensured relatively easily using extremely simple techniques, sometimes as simple as applying post-hoc filtering to the outputs of a complex model. Where the focus is placed on outcomes rather than best practices, in other words, it becomes possible to regulate AI models in a way that can be highly agnostic to the methods and data used to build the model itself. By contrast, I am skeptical that identifying "best practices" for training models can in fact generate robust consensus that a certain model is (for instance) "unbiased" if there is no robust consensus on how an unbiased model ought to behave in practice.

The implications of this argument for the regulation of AI systems are significant. Primarily, it implies that regulators should in many cases focus less on requiring AI developers to abide by certain practices, and should instead focus directly on stipulating acceptable outcomes for deployed models. In some ways, this approach can come with better regulatory traction: for instance, a regulatory approach that is strictly focused on outcomes can sidestep the need for regulators to exhaustively survey datasets or assess algorithms for "fairness," actions that would be both time-consuming and that would raise concerns about exposing intellectual property to prying eyes. However, directly stipulating what "unbiased" outcomes look like raises a separate set of thorny issues for regulators. For instance, I suggest that when discussing fairness in predictive policing algorithms, there is no model design that can give rise to a certifiably "fair" model, unless law or regulation has explicitly defined the concept of a "fair model" with reference to specific metrics that can be used to assess whether the outputs of a model are fair. But creating this definition requires regulators to define "fair" policing directly via the impacts that policing practices generate, rather than indirectly defining "fair" policing by prohibiting or promoting certain practices without regard for their aggregate impact. This is an uncomfortable task to take on, and one that likely cannot be done in a way that fully encapsulates various intuitions about what the impacts of "fair" policing ought to be.

While this argument seems to work intuitively for some specific example domains (such as predictive policing in some contexts), it is unclear if there is a general rule for when impact-based regulation is both feasible and preferable to practice-based regulation. I certainly do not believe that impact-based regulation is always feasible for all machine learning models. For it to be possible, the major requirements seem to be something like the following: 1) there is an inchoate value which is nonetheless felt to be an important restriction on how machine learning models behave; 2) there are certain cases where people broadly agree that the said value is violated; 3) a rough classification rule can be identified to identify those cases wherein the value is broadly agreed to be violated; and 4) outputs of a model can, in aggregate, be directly compared against this classification rule.

Importantly, I do not suggest that an important value needs to be reducible to a simple classification rule for impact-based regulation to be feasible, but only that it must be possible to articulate a rule that can roughly track said value. For instance, in the case of medical referrals, it is possible to imagine a large number of possible classification schemes that would fit these four criteria, e.g. "no racial subgroup should be more than 10% less likely to be referred to a doctor than any other racial subgroup upon reporting the same degree of pain on a 10-point scale." In other domains, however, it is unclear whether such measures can be formulated in principle; for example, it is not clear how the notion of "non-sexist language" could be formulated to apply across all language situations in such a way as to determine whether the outputs of a language model were effectively "sexist" or not.

The goals of this talk are to present the distinction between practice-based regulations of AI models and outcome-based regulations and to articulate how these two different frameworks can give rise to very different regulatory regimes. By focusing on a few specific domains—such as medical referrals and predictive policing—I hope to argue that the impact model of regulation is better suited for AI regulation in at least some cases. My talk will also attempt to outline the types of conditions under which this model of regulation seems plausible. If such conditions can be articulated, policymakers could be provided with a powerful set of tools to identify what types of regulatory regimes are appropriate for different types of AI models. Rather than simply being able to defer regulation until adequate technical solutions to the problems of bias, unfairness, lack of transparency, and so forth are identified, policymakers could be presented with concrete cases where immediately actionable regulations might be perfectly possible.

Brian Tebbitt (University of Minnesota, Cognitive Science)

Value Frames and the Godlike Position - Another Look at Machine Metaethics

Much continues to be written on the topic of ethics and artificial intelligence (AI), most of which addresses broad field-specific questions, as well as practical issues involved in the implementation of moral theories in machine agents, such as computability. Thus far, however, comparatively little has been written regarding the relevance of metaethical theorizing to the design, construction, and integration of artificial moral agents (AMAs). Previous work in what has been called "machine metaethics", such as that by Anderson and Lokhorst, is focused on first-order (or, practical-ethical) concerns, essentially relating to the "meta" of "metaethics" as something like one-step removal from ethics, i.e. ethical reasoning about ethical reasoning, which is not what the term is meant to imply.

Despite the importance of taking a step back to think structurally about machine ethics itself, this analysis has been more or less confined to the prescriptive (or, normative) realm. An approach to machine metaethics that bypasses nearly all second-order (descriptive) issues and insufficiently addresses questions regarding the essential nature of human morality, leaves us without a key underlying orientation vis-a-vis the central problems of alignment, control, and robot rights. A one-step removal approach skips a crucial step in the determination of our fundamental orientation toward AI and the problems of machine ethics. To fill this gap, machine metaethics must shift its focus from first-order concerns (e.g. "Are the three laws of robotics ethically suitable principles for AMAs?") to second-order concerns (e.g. "What is a moral principle, and how is one derived?"), or at least address them in its overall analysis.

Shifting from first to second-order concerns involves, I would like to suggest, two "reckonings": [1] a reckoning with what morality is, and [2] (pursuant to [1]) a reckoning with the nature of human values, or human valuation. Beginning with the formulation of Stich, I consider whether morality is an "elegant machine" (rationalistic, idealistic, and involving a consistent, top-down logic) or a "kludge" (bottom-up, arising from a collection of psychological mechanisms against the backdrop of culture) and determine that it is indeed kludge-like and relative to values. The elegant machine hypothesis is set aside as descriptively uncompelling, though it is favored by many machine ethicists and computer scientists because such a conceptualization is favorable to computability.

Since morality is, descriptively speaking, a function of human valuation, we need to reckon with the nature of human valuation, and valuation generally—regardless of the agent under discussion. To this end, I introduce the novel concept of value frames, a cognitive scientific notion based in embodiment. The idea is that each agent, and each kind or species of agent, occupies their respective value frame, which is the set of what that agent (or species of agent) tends to value. And since agents cannot be embodied other than how they are, they are unable to transcend their particular value frame. Thus, for the agents that occupy them, value frames are inescapable. And this is the outcome of reckoning with what morality is and its relativity to values.

A failure to reckon with the nature of morality and values, along with a popular sense that AI researchers are reaching beyond themselves to create autonomous agents, leads to a godlike position. The godlike position is a way of viewing the construction and development of AI and machine agents such that one is misled into thinking that it is possible for us to transcend the human value frame while still (at the same time) being concerned with controlling our creations and ensuring that machine agents do what we want them to do, or refrain from doing what we don't want them to do. In this position, we view ourselves as "gods"to machine agents, asserting our unbounded will, rather than seeing the truth that, for better or worse, machine agents are an extension of our own values.

The question of value frames, in full acknowledgement of the nature and source of human morality, rightly leads us to consider a fundamental question (and perhaps the fundamental question) of machine ethics: Should we pursue imitation or self-determination with regard to machine agents, and especially with regard to any future AMAs? Imitation means programming machine agents to imitate human moral decision-making, and to act in accordance with what humans view as ethically or socially acceptable. Self-determination, on the other hand, means creating machine agents that will develop their own values and goals as a sort of "species" unto themselves. Under the assumption of imitation, programming and design occurs squarely within the human value frame, while under self-determination AMAs transcend the human value frame by being allowed to develop and maintain their own.

The pre-ethical choice between imitation and self-determination, between the human value frame and the possible development of another (non-human) value frame is pivotal. Selecting either will largely determine our basic stance toward alignment, control, and robot rights. Choosing imitation effectively means requiring alignment and control, and possibly rights for AMAs. Choosing self-determination, however, means relinquishing certain claims to control and alignment (and maybe all of them).

Drawing broadly on the work of machine ethicists (e.g. Anderson, Allen), computer scientists and researchers (e.g. Russell, Christian), and philosophers (e.g. Nietzsche, Bunge, Stich), I take another look at machine metaethics and propose a fresh way of applying metaethical theorizing to machine ethics, and to AI generally.

Matt Stichter (Washington State University, Philosophy)

Flourishing Goals, Metacognitive Skills, and the Virtue of Wisdom – Putting Emotion and Moral Motivation into Place

In this talk, I'll elaborate some details of an account of wisdom based on my philosophical account of virtue as skill, and its grounding in self-regulatory and goal-oriented theories in psychology and cognitive science. I'll further discuss how this compares and contrasts to both the common wisdom model discussed by Grossmann (CWM), and a rival account of wisdom based on Aristotelian phronesis proposed by Kristjánsson (APM). I'll focus on resolving two areas of contention between those models: accounting for moral aspirations and motivation (by drawing on goal theory); and the role of emotion in wisdom (by suggesting an enactivist approach).

John Sullins (Sonoma State University, Philosophy)

Machine Wisdom, Artificial Phronesis, and Robot Inner Dialog

Inner speech is a concept from psychology that suggests that the inner dialog many of us experience as we accomplish tasks helps us become conscious of our thoughts or bring to consciousness the salient aspects of the problem at hand. This inner dialog also plays a role in skilled moral and ethical reasoning. Robot inner dialog has been used to build systems that display more conscious and trustworthy actions. In this paper we explore the possibility of testing aspects of artificial phronesis, or skilled practical moral reasoning, in machines through extending work done on the development of robot inner speech.

Michael Tamir (University of California Berkeley, School of Information) & Elay Shech (Auburn University, Philosophy)

Understanding and Deep Learning Representation

Advances in Machine Learning (ML), especially using Deep Learning (DL) techniques, have accelerated performance in numerous areas of practical application. One metric worthy of attention is the rate at which DL has enabled algorithms to compete with human benchmarks on specific tasks. Image classification, for instance, has evolved dramatically thanks to a series of specific improvements in DL, including Convolutional Neural Network (CNN) architectures coupled with technical advances in the optimization of neural networks with multiple hidden layers, leading to DL beating human performance on the ImageNet benchmark dataset (He et al., 2015). AlphaGo's defeat of Lee Sedol in 2016 is another celebrated example of DL in interactive reinforcement learning contexts. Similarly, better than human deep reinforcement learning successes were achieved by OpenAI in Dota2 competitions, and CMU's Libratus and Plaribus poker algorithms. More recently, tasks in modern natural language processing (NLP) have also seen ostensible breakthroughs by becoming competitive with human performance. Hassan et al. (2018) achieved parity with human translation on the WMT17 benchmark, leveraging DL Transformer architectures. Transformer architectures have also inspired a wave of advances leading to performance increases on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al. 2018), overtaking non-expert human performance in Nangia et al. (2019). Similarly, over a dozen DL Transformer based techniques currently beat human performance scores on the Stanford Question Answering Dataset 2.0 (SQUAD 2.0) (Rajpurkar et al. 2018).

Human competitive performance on such benchmarks has accompanied an increased use of terms like "understanding" in artificial contexts. Machine understanding of natural language applications is commonly discussed by researchers both in terms of task goals as well as model capabilities. The GLUE benchmarks in "Natural Language Understanding" (NLU) tasks are framed in terms of "aspir[ing] to develop models with understanding beyond the detection of superficial correspondences between inputs and outputs" (Wang et al. 2018). The SuperGLUE benchmark lists as the first criteria that "[t]asks should test a system's ability to understand and reason about texts" (Wang et al. 2019). Devlin et al. (2019) motivate specific techniques "[i]n order to train a model that understands sentence relationships," while Raffel et al. (2019) more modestly claim that techniques such as those of (Devlin et al. 2019) "can be loosely viewed as developing general-purpose knowledge that allows the model to 'understand' text." Researcher discussion of machine understanding is even bolder in areas focused on DL representation learning. Bengio et al. (2013b) influentially framed conversations on machine understanding in terms of disentanglement, arguing "the ultimate goal of AI research is to build machines that can understand the world around us, i.e., disentangle the factors and causes it involves." Chen et al. (2016) motivate using generative techniques with "the belief that the ability to synthesize, or 'create' the observed data entails some form of understanding, and it is hoped that a good generative model will automatically learn a disentangled representation," and Higgins et al.

(2017) claim that representations with disentangled factors are "an important precursor for the development of artificial intelligence that understands the world in the same way that humans do."

DL successes coupled with such loose (if not bold) claims about potential machine understanding have prompted responses from intersecting research in cognitive psychology (Marcus 2020, Lake et al. 2017) and linguistics (Bender et al. 2020). These responses make the easy case that models trained for human competitive performance in specific tasks fail to possess what Marcus calls "deep understanding" (like that found in humans), citing failures to perform when "circumstances deviate from training data" (Marcus 2020). While few claim that current algorithms possess such "deep" or "human level" understanding, the more interesting question of which conceptual criteria are appropriate for evaluating (partial) machine understanding has not received sufficient attention. Can the philosophy of science and epistemology literature on understanding shed light on which conceptual criteria are important for machine understanding? Are there trends and patterns in how DL trained algorithms process data from a representation and information compression standpoint that could partially or fully satisfy such conceptual criteria? If so, do such patterns provide insight into critically evaluating and interpreting the relevance of concepts like "understanding" in an artificial context?

In this work, we answer these questions, identifying three key factors taken from the philosophy of understanding literature which we argue have a basis for evaluation in DL trained algorithm performance and learned data representations. Our aim is twofold. First, viewing DL successes in the context of philosophy of understanding may shed light on the extent to which references to "understanding" in DL research have any grounding (or not) in traditional analyses of the concept. Second, the philosophy of understanding literature provides valuable resources for identifying the conceptual criteria that are appropriate for evaluating partial applicability of concepts like "understanding" to machines. Using these resources, we identify methods for experimentally detecting the presence of key factors indicative of understanding, allowing for future evaluation of potentially more complete or so called "deep" machine understanding in the rapidly evolving field.

We lay out the paper as follows. Drawing from select philosophical accounts of understanding in Section 2, we identify reliable and robust task performance, as well as information relevance and well-structured representation, as three key factors. In Sections 3 we provide a brief introduction to ML and DL practices. While even successful individual DL trained algorithms are not minded agents, in Section 4 we show how phenomena analogous to said factors can be observed and evaluated in DL applications through an information theoretic analysis. Specifically, deep neural networks use multiple layers of representations that systematically learn to extract and organize relevant information, and this process directly relates to methodologies used by DL researchers to ensure reliable and robust success. Information relevance is directly learned by the neural network, preserving task-relevant information in deeper hidden layer representations of the raw data. When successful, learned representations

develop insensitivity to unimportant factors while optimally leveraging and organizing the relevant features, thereby disentangling raw details in deeper layers based on (task) significance. Section 5 ends the paper with a consideration of two related objections to evaluating "understanding" in the context of automated task performance. Our goal is to establish a discussion framework for understanding in ML and DL, and to encourage future investigation grounded in philosophically coherent terms and direct engagement with the technology driving these accomplishments.

Shannon Vallor (University of Edinburgh, Philosophy)

Moral Mirrors and Telescopes: The Opportunity for AI-Augmented Wisdom

In this talk I review the core components of wisdom embedded in mature philosophical and psychological accounts, and their implications for the prospect of machine wisdom. I argue that certain core components of wisdom are either fundamentally inaccessible to machine agency, or superfluous to it, being meaningful only for entities like us who suffer from biological impediments to holistic, balanced and other-oriented value judgments. However, I show that other core components of wisdom present substantial opportunities for machine-mediated augmentation of human wisdom, which may enable human flourishing far more reliably than existing efforts to employ artificial intelligence for good.

Pittsburgh Blitz Session (3 short talks)

Xin Hui Yong (University of Pittsburgh, Philosophy)

A Seat At The Table? Modelling Whether Algorithmic Agents Compound the Dampening of Minoritized Voices

In highly unequal societies, where there are clear demarcations between those with higher privilege and power and those without, agents at the top of the social hierarchy (call them elite agents) are often ignorant of the struggles of those at the bottom. From a policy and decision-making front, these same agents with higher privilege and power are disproportionately placed in positions of societal decision-influencing and making power (such as economists, lawmakers and political leaders), and are thus less likely to be equipped with knowledge about non-elite groups.

Often through work of these minoritized groups, there has been more awareness of the harmful effects of elite agents being the only ones to make the decisions. Therefore, given the recent increased salience of the importance of letting minoritized voices speak and amplifying their voices, many institutions have made efforts to include underrepresented members of the community in important decisions regarding that community. This would be in line with standpoint epistemology as well as accounts of silencing; the hope is that if we give the minoritized "a seat at a table," they will finally be able to speak to their experiences and influence decision-making in favour of the already disenfranchised. At the same time, though, and perhaps even due to discussions about reducing human bias, sociopolitically relevant decisions, such as recidivism or housing decisions, are made with the consultation of algorithmic tools. Many of these algorithmic tools are trained with past human decisions, both as training data and as a benchmark, and thus could inherit any errors or biases in these historical decisions.

Given the above, I have two worries: The first (I) is that the individuals who are included because of their minority background would have less of an impact in eventual decisions because of their recency. In other words, the weight of this new seat at the table reduces the longer the delay of the addition, compared to the weights of the other seats in terms of this new addition's influence on the final decision. The second (II) is that if the disparity of impact in (I) is an injustice, then the use of algorithmic tools that blithely rely on historical decisions will only compound this injustice by increasing the inequity in influence. To illustrate my worries, I model a panel of agents that aim to converge on a decision. I aim to reflect the phenomenon where many decision-making panels have consciously included minoritized members, but also use algorithmic tools to aid them with their decision-making process. I adapt Kevin Zollman's social epistemology model from Chapter 4 of Network Epistemology(under contract), in which each epistemic agent 'pools' their current guess of a number (say, how much to weigh SAT scores in undergrad school application packets) with other epistemic agents they are connected

to, with the aim of getting the best estimate. The 'pooling' is such that the agents will revise their guesses to be closer to their peers', eventually converging with them.

I tweak the model in two ways:

1. I compare two cases, where a. one agent is added in and able to pool with the other agents after n iterations, and b. the agent, with the same initial guess, was added in at the first iteration of the pooling. The aim here is to show (I) above. Case a. models the 'late' inclusion of the seat at the table, with b. as a control setup where the beliefs of the agents are all the same as before, but all agents are able to pool with each other from the start.

Once the agents are connected, the model does not discriminate between the latecomer agent and the other human agents – once any human agent is incorporated into the network, they are treated the same by the other agents. This is both to reduce the complexity of the model, and also an attempt to model how some policies only encourage inclusion of minoritized voices, but do not privilege these voices once they are included (perhaps for fear of being biased). As a future project, I welcome suggestions on how best to model the further amplification of these minoritized voices.

2. As for (II), I compare case 1b with case 2, where I add in one or more 'algorithmic' nodes that aggregate all historical and current guesses, and I have each 'human' epistemic agent also pool over these 'algorithmic' nodes as if they were also fellow human epistemic agents. Just like in case 1., 2a. models a 'late' inclusion of a human agent, whereas in 2b. all the agents are included in the complete network from the first iteration.

Note that this should absolutely not be seen as an accurate representation of human-machine interactions, or even of the algorithms that would be employed in a decision-making panel such as an admissions committee. As the model grows in sophistication from its baseline, I'd love to hear more about empirical findings regarding the efficacy of these equity boosting measures so I can better incorporate them into the model! In addition, I recognize that I am starting with as little moving parts as possible, and welcome suggestions on improvements in the modelling of these algorithmic tools.

Through some preliminary simulations with a complete network of agents (with the late agent joining the complete network on the 51st iteration in cases 1a. and 2a.), I have found that the late inclusion of the agent increases the difference (call this d) between the eventual belief that the group converges on and the latecomer's initial belief, compared to if they were included at the start. As for the impact of the addition of the 'algorithmic' node, it depends on how many iterations the other agents have had before the new agent is added, and how many iterations are required before convergence. With more iterations before the latecomer's arrival, the addition of the algorithmic agent increases d, but if the agent is added in earlier, d is reduced. For example, d in 2a is actually reduced compared to 1a., and the time to convergence is also decreased.

I hope that through presenting the models and my findings at the Machine Wisdom Workshop, I will invite further discussion on how the timing of an intervention (be it algorithmic or otherwise) can affect its efficacy. I also hope this model could go into furthering arguments that representation and inclusion of minoritized individuals, in itself, is not a sufficient solution to long-standing inequities perpetuated when the ignorant elite make decisions for the masses. I'd also love to learn more about different harm-reducing interventions in human-machine interactions, and how they could be incorporated into models such as these. In addition, I would be excited to hear from experts in other fields about empirical research on algorithms and inclusion, as well as how they could corroborate, disagree or be incorporated into my models.

Conny Knieling (University of Pittsburgh, Philosophy)

Arbitrariness in Algorithmic Decision-Making as a Moral Problem

Decisions are everywhere and there are many moral problems that arise when it comes to decisions. Arbitrariness is one of the lenses through which one can judge decisions and decisions being arbitrary is not unproblematic from an ethical perspective for the people affected by said decisions. When it now comes to the increasing use of automated decision-making involved in or replacing human decision-making, the question of whether and in what extent arbitrariness by automated decision-making systems is of moral concern naturally comes up.

One possible ethical examination of the problem of arbitrary decisions made by algorithms was offered by Creel and Hellman (2021) and they concluded that arbitrariness except in special cases or when applied across a broad range of cases does not pose a problem. In this talk, I will argue the contrary, that even isolated arbitrary decisions can wrong the individual who is affected by this decision and that this harm is constituted qua being arbitrary. For this, I will first provide a positive proposal about what "arbitrariness" might amount to in decision-making broadly, and in algorithmic decision-making in particular, that goes beyond what Creel and Hellman (2021), among others, have identified.

Second, I will concede that not all arbitrary decision-making needs to be of ethical concern and that the label of arbitrariness might not be applicable for many decisions. Nevertheless, when arbitrariness becomes a moral issue, it constitutes a harm for the individual affected qua the decision being arbitrary and this harm persists even when the automated decision-making systems can be considered “fair”. I will argue here that the literature on fairness has overlooked the importance of procedural considerations in algorithmic decision-making and that looking at the political philosophy literature can help us inform this question. It is the idea of “due process” and procedural justice that can bridge a persistent gap that has not yet been addressed in the literature on fair AI nor could it be addressed within the framework of “fairness”. Given all this, I will finally provide an ethical evaluation of how we can understand arbitrary decision-making from a moral point of view and what this means for algorithmic decision-making.

While this talk in its motivation is a reply to Creel and Hellman (2021), the claims provided will go beyond the scope of their paper, and a main goal of my talk will be to bring research on procedural injustice and “due process” in conversation with the research on fair algorithms and explanatory AI.

Konrad Werner (Center Visiting Fellow, University of Warsaw, Philosophy)

The Machine Wisdom of Not Being Too Wise: Social Apps and Cognitive Confinement in the Time of Mental Health Crisis

In late January 2022 *Politico* revealed morally questionable data management practices in Crisis Text Line. This is a company described as "one of the world's most prominent mental health support lines, a tech-driven nonprofit that uses big data and artificial intelligence to help people cope with traumas such as self-harm, emotional abuse and thoughts of suicide." I turns out that the company shares anonymized data with a for-profit company named Loris, creating software for customer support services. Its goal is to make AI customer assistance more sensitive to individual needs. The anonymized data packages they get from Crisis Text Line are supposed to make the conceptual framework employed in these AI customer serviced more fine-grained.

Another example I shall refer to – this time not involving any scandalous activities – is a Polish startup called Mindgram. It's a social app, as its website informs, "where you can improve your mental well-being, learn how to manage your thoughts and emotions, support your body, or build your support system through relationships with people." So, it's a broader on-line health and wellbeing service.

Aside from privacy, the listed software tools bring the danger of medicalization of mental health and wellbeing. There is also a correlated problem of incentives on the side of the companies to not only answer demand but also create demand. But these are things I shall discuss.

In my presentation I want to approach the listed cases in cognitive/epistemic terms:

1. "Cognitive niche" is a very general term referring to the domain determined by a subject's cognitive capacities; an environment in which the subject has access to specific information resources. Metaphorically speaking – X's cognitive niche *opens the world* to X.

2. This general concept can be specified and operationalized in many ways. In my talk I shall cash it out in terms of so-called epistemic dependencies. The latter in turn refers to epistemically relevant relations between individuals in a certain community, determining what the individuals know and – crucially for my project – what the individuals problematize, thus what questions they ask.

3. "Cognitive confinement" refers to a pathological form of the cognitive niche, making it impossible for the subject to get access to specific information resources. In the talk it will also be articulated in terms of epistemic dependencies.

4. Some online communities can be theoretically approached as cognitive niches – in this case I will use the term "virtual cognitive niche" which already came up in the relevant literature. Accordingly I will also speak of virtual cognitive confinements.

5. Online filter bubbles can be represented as virtual cognitive confinements, which is a view I argued for elsewhere. In this talk I will try to represent the two examples given earlier – the Crisis Text Line + Loris case and the Mindgram case – in terms of virtual cognitive niche/confinements.

6. Here is the claim I shall be arguing against: As part of a bigger effort to create a human-AI collaborative "unities" based on practical wisdom we should support initiatives such as Loris and Mindgram. For, according to the argument, we should be able to create AI solutions that are better suited to human needs, including mental, emotional and moral needs. I this spirit, when we open Loris' website we see their slogan which says "Using machine learning to make customer support more human, empathetic and scalable."

7. I define "machine practical wisdom" in the context under investigation here as a set of practices, led by some operationalizable beliefs, that are dedicated to preventing virtual cognitive niches to pathologize and morph into virtual cognitive confinements.

8. With this characterization of machine practical wisdom, someone might rearticulate 6 in the following way: a virtual cognitive niche becomes a virtual cognitive confinement when the fine-graininess or resolution of questions and answers people process there decreases; in simple terms – when their world-view becomes less sensitive to details and context. Therefore, making mental health and wellbeing AI more fine-grained, thus increasing its resolution, is a means to improve practical machine wisdom.

9. I shall argue that the opposite is true in this particular case – it's quite peculiar at first glance, indeed – namely that less fine-grained (to a degree, of course) questions and answers serve practical wisdom better in the AI context. I oppose two moral categories, re-defined in "machine" terms: sympathy (which in used in Loris' slogan) and mercy, and argue in favor of "machine" mercy and against "machine" sympathy.

10. The crucial difference between the two – from the perspective undertaken in the talk, and of course given a certain degree of conceptual engineering involved here – is that there is no inherent limit to sympathy in terms of how fine-grained the questions penetrating our mental conditions can be. To put it in a somewhat simplistic way: for any well-being problem P such that a solution to it seems to be attainable, there is a new well-being problem R that is potentially implied by P, potentially implying P, or elicitable inside P once P gets analyzed, such that finding a solution to R requires further effort (thus further AI support). Meanwhile – let me stress again, given a certain degree of conceptual engineering – there is a built-in limit to mercy because of the act of forgiveness included in it. Which refers to forgiving others but also forgiving oneself. This act also requires a certain fine-grained analysis of what is supposed to be forgiven, but its goal is not analysis, but an act of moral absolution dedicated to establishment new grounds for future actions.

11. The new ground brought forth by forgiveness is a theoretical fiction, but it plays – it can be argued based on literature in ethics and moral psychology – an important role in moral life as well as in contributing to functional societies. Therefore, it is a crucial ingredient of practical wisdom. This should include machine practical wisdom referring to human-AI interactions. Therefore I am going to speculate a bit about practical (here – institutional, which is my major concentration) means to impose certain resolution limits, as it can be spelled out, on the degrees to which mental health and wellbeing software should be allowed to penetrate our lives. For it is not so wise to be too "wise" – in the sense implicitly promoted e.g., by Loris, which stands for being "sympathetic."