ABSTRACTS

COPENHAGEN WORKSHOP ON ALGORITHMIC FAIRNESS

NOVEMBER 12-13, 2020

Algorithms, Fairness, and Social Justice (Ben Eidelson, Harvard Law School)

Much of the discussion of "algorithmic fairness" focuses on the risk of error and on how that risk is distributed. This kind of concern is exemplified by the objections to recidivism models that yield more "false positives" for black defendants and more "false negatives" for white ones. But the use of algorithmic predictions also raises a distinct set of concerns that do not depend on the risk of error at all—indeed, that gain force in proportion to the algorithms' accuracy. If whatever counts as merit for some purpose is unjustly distributed, then a decision procedure that accurately identifies merit and differentiates on that basis will "pick up" the pre-existing injustice, and thereby potentially extend or aggravate it, in ways that more random, less merit-tracking selection processes would not. In this essay, I draw on theoretical accounts of the normative foundations of discrimination law to argue for the centrality of this latter kind of problem and to critically evaluate different ways of understanding its contours. I distinguish between two types of concerns: first, that an allocative decision may wrong a person by "compounding" a prior injustice that she, individually, suffered (as Deborah Hellman has argued); and second, that decision procedures may contribute to future social injustice by sustaining or aggravating patterns that undermine equality of status and opportunity. I raise doubts about the first idea and argue for the importance of the second. In normative assessments and legal regulation of algorithmic decision making, a central concern—perhaps the central concern—ought to be the potential for this practice to entrench harmful and unjust patterns, quite apart from any alleged unfairness or other personal wrong to the individuals about whom the predictions are made.

Algorithmic Discrimination and Compounding Injustice (Kasper Lippert-Rasmussen, Aarhus University)

In the context of differential treatment based on the use of algorithms (and for that matter in relation to indirect discrimination, more generally), it is sometimes objected that the relevant kind of differential treatment compounds injustice and that we have a right that others do not compound injustices to which they have been subjected (Hellman 2018; Hellman 2020). Roughly, one compounds injustice when one appeals to facts that obtain as a result of prior injustice to justify imposing what in effect are further disadvantages on the victims of prior injustice. For instance, if women are unjustly forced or pressured into taking more responsibility for childcare than men, e.g., by taking longer periods of parental leave in connection with childbirth, and employers then use the statistical fact that women take more parental leave as a reason for giving preference to male applicants, then employers compound the injustice against women. In this paper, I first discuss what compounding injustice amounts to in the context of algorithmic discrimination providing some conceptual and normative complexity in addition to that which has already been uncovered in the literature. Next and based on the conceptual and normative groundwork in the first part of the article, I critically assess the view that we have a duty not to compound injustice in the context of algorithmically based discrimination. Generally, I am skeptical of the existence of such a duty not to compound injustice (as opposed to an extensionally largely overlapping duty not to make anyone who is already unjustly off even worse off). I do agree, however, that many of the cases that are analyzed as involving a violation of a duty not to compound injustice do involve violation of some moral duty or other. Also, I try to answer the question of what the most plausible version of such a duty is assuming that we have one.

The Algorithmic Leviathan: Arbitrariness, Unfairness, and Opportunity in Algorithmic Decision Making Systems (Deborah Hellmann & Kathleen Creel, University of Virginia School of Law)

Automated decision-making systems implemented in public life are typically standardized. One algorithmic decision-making system can replace thousands of human deciders. Each of the humans so replaced had her own decision-making criteria: some good, some bad, and some arbitrary. Decision-making based on arbitrary criteria is legal in some contexts (employment) and not in others (criminal sentencing). Where no other right provides a guarantee of non-arbitrary decision-making, is arbitrariness of moral concern? We argue that an isolated arbitrary decision need not morally wrong the individual whom it misclassifies. However, if the same algorithms are applied across a public sphere, such as hiring or lending, the same people could be consistently excluded. This harm persists even when the automated decision-making systems are "fair" on standard metrics of fairness. We argue that such arbitrariness at scale is morally problematic and propose technically informed solutions that can lessen the impact of algorithms at scale and so mitigate or avoid the moral harms we identify.

Fair Equality of Chances for Statistical Prediction-Based Decision Making (Michele Loi, University of Zürich, Hoda Heidari, Cornell University, and Anders Herlitz, Institute for Futures Studies, Stockholm)

Our paper presents a fairness principle that can be used to evaluate decision making based on predictions. it characterizes – in a formal way – how luck must impact outcome in order for its influence to be considered fair. The framework can be used to evaluate rules of decision making on the basis of different moral theories, and is compatible with the broadest range of moral views according to which inequalities due to brute luck can be fair.

We propose that a decision rule based on predictions is fair when the individuals directly subjected to the rule enjoy fair equality of chances. We define fair equality of chances to obtain if and only if the individuals who are equal with respect to the features that justify outcomes have the same statistical prospects of being benefited or harmed by the decision rule, irrespective of their morally irrelevant traits. We show that fair equality of chances corresponds to one statistical fairness criterion (sufficiency) in some circumstances and to another statistical fairness criterion (separation) in other circumstances, depending on whether decisions and actual outcomes are identified as the benefits or justifiers of inequality for the individuals affected by algorithmic decisions. We provide a mathematical proof of this claim in the appendix, and we argue for it in the main text of this article. The paper is structured as follows. In (2) we illustrate the problem with a simple hypothetical example of a predictor of drunk driving applied to two populations, Christians and Muslims, who differ relative to their base rate of drunk driving and to the incidence of the features that are used to make the prediction. Then, we introduce an example of prediction-based decision making. This example also illustrates the tension between different statistical fairness criteria. These tensions have led some to believe that it is impossible for prediction-based decision making to be fair. In section (3), we sharpen the dilemma by introducing and explaining conventional fairness criteria used in statistics. In section (4) we argue that outcomes due to brute luck sometimes can be fair. In Section (5), we introduce our proposal: fair equality of chances. In section (6) we show that two of the most widely discussed statistical fairness criteria, separation and sufficiency, correspond to mutually incompatible interpretations of the principle of fair equality of chances. There is a brief concluding section.

Fairness, Classification Parity, and the Levelling Down Objection (Sune Holm, University of Copenhagen)

What does it mean for algorithmic classifications to be fair to different groups? Classification parity definitions require equality across groups with respect to some algorithmic performance measure such as error rates or predictive values. The talk first presents a philosophical argument for classification parity based on an egalitarian argument schema. Focusing on the equality of opportunity definition of algorithmic fairness proposed by Hardt et al. (2016) the talk presents an instance of the levelling down objection to that definition: It seems absurd if our definition of algorithmic fairness means that a fair algorithm will sometimes be worse for some and better for no one than an alternative. Finally the talk assesses the merits of two responses that proponents of classification parity might present to the levelling down objection.

Do Groups Matter in Algorithmic Discrimination? (Shalom Chalson, ANU)

Discrimination refers to acts, practices, or policies which distinguish between people, treat them differently, and disadvantages them. As a normative concept, it captures the intuition that it is wrong to hold people captive to protected attributes, features which signify one's membership of a social group —like one's race, religion, sex, or gender—,as well as to treat them unfavourably in light of their possessing those features. Discrimination is taken to occur to individuals qua members of social groups. Some philosophers have paid particular attention to the role of social groups in describing the wrongfulness of discrimination (Pincus 1996; Lippert-Rasmussen 2006, 2011; Arneson 2006, 2013), while others have suggested that an appeal to groups is unnecessary (Eidelson 2015; Moreau 2020). In this paper, I consider whether we must make reference to groups in order to adequately describe the wrongfulness of discrimination. In order to do so, I examine the case of algorithmic discrimination. With the increasing use of models built from algorithms in domains like healthcare and criminal justice, algorithms may perpetuate unjust social structures or create new paradigms of unfairness and must be considered more closely. For some theorists, it is members of socially salient groups, or groups which structure interactions across a variety of contexts (Lippert-Rasmussen 2006, 2019), who experience the harm of discrimination. On this conception, ML is merely a new site of discrimination that perpetuates injustice by naming and harming groups explicitly. In my view, however, this fails to capture a class of cases I deem paradigmatically discriminatory: disparate treatment where discriminatees are not members of socially salient groups. ML, in particular, may easily produce unjust outcomes by making use of proxies for group membership, firstly, and creating new groups which then become recipients of systematic injustice, secondly. I hold that wrongfully discriminatory acts share a structure; X discriminates against Y when X subjects Y to (i) differential treatment that (ii) unfairly disadvantages Y. Further, the difference-making feature that accounts for why Y and not Z, Y's counterfactual other, was treated as such is a property that is socially salient. It is unfair disadvantage, and not the fact that discriminatees are members of social groups, that motivates the claim that discrimination is wrongful. But it is disadvantage that gets at the importance of groups too: (a) socially salient groups are likely to illuminate the features which are themselves socially salient (clumps of properties with meaning tacked on to them) and (b) these groups are likelier targets of unfair disadvantage and, accordingly, harm, than random clusters of individuals. Crucially, social groups are not a necessary condition for discrimination.

Counterfactual Fairness as Machine Fairness? (Clinton Castro Florida International University, David O'Brien (Tulane), and Ben Schwan (Case Western Reserve University)
Predictive analytics influence consequential decisions in nearly all facets of modern life. This has given rise to the young field of fair machine learning and a number of fairness measures, mathematically circumspect definitions of fairness that purport to determine whether a given predictive system is fair. Following Binns (2018), we take "fairness" in this context to be a placeholder for a variety of normative egalitarian considerations. We explore a few fairness measures to suss out their egalitarian roots and evaluate them, both as formalizations of egalitarian ideas and as assertions of what fairness demands of predictive systems. We pay special attention to a recent and popular fairness measure, counterfactual fairness, which holds that a prediction about an individual is fair if it is the same in the actual world and a counterfactual world where the individual belongs to a different demographic group (cf. Kusner et al. 2018).

Second-order Theorizing in Algorithmic Fairness (Otto Sahlgren, Tampere University)
Fairness, (in)equality and non-discrimination have become central issues in the ethics of AI. To address issues with algorithmic discrimination and unfairness, the fair machine learning community has introduced a plethora of definitions for 'fair algorithms' as well as technical methods for mitigating bias (see Verma & Rubin 2018). Legal and philosophical notions of fairness, equality and (non-)discrimination translated into benchmarking metrics for machine learning models allow model developers to "measure" fairness in terms of fairness of treatment (i.e., whether the system treats individuals fairly) or impact (i.e., whether the generated probability scores and classifications are fair across groups). What can be called "first-order theorizing" in algorithmic fairness – i.e., efforts to construct universally applicable definitions and methodologies for fairness in algorithms — has been subjected to significant criticism, however. Critical voices emphasize the inherently social, political and contestable nature of fairness (Narayanan 2018; Green & Hu 2018). Current first-order approaches, critics argue, are 'ideal' (in the pejorative sense) and limited in their scope, neglecting individual rights and existing inequalities (Fazelpour & Lipton 2020; Herington 2020). Further arguments point to their insensitivity to social and contextual factors that give meaning to fairness claims (Selbst et al. 2019), including relevant "currencies" of equality that are at stake (Binns 2018). The paper contributes to research in algorithmic fairness in a two-fold manner. First, it analyzes the ostensible shift in the fair machine

learning literature towards what is called second-order theorizing about algorithmic fairness. Second-order theorizing focuses on dealing with the (sometimes conflicting) claims of substantive first-order theories. Allowing for reasonable disagreement about fairness, second-order theories aim to explicate conditions for and/or outline procedures that result in correct and/or justified decisions regarding fairness and performance in algorithms even at the face of trade-offs. Approaches under this heading vary significantly, from participatory policy-design in algorithms (Lee et al. 2019) to deliberative democratic models for algorithmic fairness (Wong 2019). The paper examines some prominent second-order approaches, articulating distinctive features of second-order theorizing in algorithmic fairness. Secondly, drawing on insights from political and critical theory, the paper suggests some future directions and critical questions for second-order theorizing about algorithmic fairness. It is argued that, in order to "practice what it preaches", second- order theory must be self-reflexive and critical: A robust second-order theory should, firstly, articulate and justify its normative and theoretical commitments and, secondly, seek to identify barriers for application of the theory in practice.

On Algorithmic Fairness in Medical Practice (Thomas Grote, University of Tübingen & Geoff Keeling, Stanford University)

The application of machine learning technologies to medical practice promises to enhance the capabilities of healthcare professionals in the assessment, diagnosis, and treatment, of medical conditions. However, there is growing concern that algorithmic bias may perpetuate or exacerbate existing health inequalities. Hence it matters that we make precise the different respects in which algorithmic bias can arise in medicine, and also make clear the normative relevance of these different kinds of algorithmic bias for broader questions about justice and fairness in healthcare. In this paper, we provide the building blocks for an account of algorithmic bias and its normative relevance in medicine. The paper proceeds as follows: In part I, we give a brief outline of the applications of machine learning in medical practice and point out how issues of (un-)fairness manifest within this context. In part II, we investigate the mechanisms that are causally relevant for algorithmic discrimination. To this end, we develop a tripartite account of algorithmic bias, which distinguishes between formal, substantive, and normative notions of bias. Also discussed will be the different sources which give rise to these biases within healthcare. Whereas the first two parts provide the conceptual and empirical grounding, the remainder of this paper tries to develop a substantive account of algorithmic fairness, while also discussing possible steps to facilitate fair decision-making within medical practice. Thus, in part III, we consider different formal metrics for fair-decision-making. We argue that Deborah Hellman`s principle of 'error ratio parity' is best suited to the demands of medical practice. Ultimately, however, we argue that a merely formal account of fairness falls short in capturing the normative character of algorithmic fairness within medical practice. In this respect, part IV explores ways to develop a more comprehensive theory of fairness, first by incorporating an account of procedural fairness which is largely inspired by work on causal inference in machine learning, and second, by revisiting the normative foundations of fairness in medical practice, drawing on work from John Broome and Brad Hooker.

Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Crime Prediction (Tom Douglas & Ben Davies, Oxford University)

It is often thought that traditional crime prediction tools, though biased in many ways, can straightforwardly avoid one particularly pernicious type of bias: direct racial discrimination. They can avoid this by excluding race from the list of variables employed to predict offending. A similar approach could be taken to the design of newer, machine learning-based (ML) tools for predicting crime: information about race could be withheld from the ML tool during its training phase, ensuring that the resulting predictive model does not use race as an explicit predictor. However, if race is correlated with measured offending in the training data, the ML tool may 'learn' a perfect proxy for race. If such a proxy is found, the exclusion of race would do nothing to weaken the correlation between risk (mis)classifications and race. Is this a problem? We argue that, on some explanations of the wrongness of discrimination, it is. On these explanations, the use of an ML tool that perfectly proxies race would (likely) be more wrong than the use of a traditional tool that imperfectly proxies race. Indeed, on some views, use of a perfect proxy for race is plausibly as wrong as explicit racial profiling. We end by drawing out three implications of our arguments.

Three Dimensions of Algorithmic Fairness (Fabian Beigang, LSE)

Algorithms increasingly play a large role in deciding things about and for humans. In doing so, they must have access to relevant information about those humans lives and identities. However, there has been disagreement over what aspects of human identity are ethical to use in decision making. Multiple studies demonstrate that bearers of historically oppressed identities, including racial and gender identities, receive negatively biased results in algorithmic decisions (Noble, 2018). One of the common objections to the use of these variables is that they can generate negative feedback loops, wherein the algorithm predicts outcomes while simultaneously causally contributing to those outcomes. In this paper, however, I wish to explore a different way in which the use of these identifiers contributes to unfair outcomes. I argue that variables such as race and gender make poor variables for algorithmic decision making because they have controversial extensions. That is, it is indeterminate what people belong in the extensions of racial and gender terms. While most demographic terms are at least minimally ambiguous, some have more conventionally accepted meanings than others. For example, we know that when we are asked for age, the answer should be given in years, corresponding to the amount of time the earth takes to travel around the sun. Race and gender, however, are two variables that currently lack a stable conventionally accepted meaning. There is currently, both in philosophy and in broader society, an ongoing debate on how best to conceptualize the meanings of these terms. These debates have moved beyond the academic sphere into the public realm, where they have practical, real world consequences. The varying conceptualization of the meanings of race and gender mean that the same individual might, on different accounts of race and gender, be categorized as 'male', 'female', 'white', 'mixed', 'black' or none of the above. The controversial nature of variables such as race and gender leads to several potential problems for algorithmic decision-making. In this talk, I will discuss two problems associated with controversial extensions. The first problem is that the controversial extensions of race and gender can negatively influence data quality. An individual's self-reported gender/race may not correspond to the algorithm developer's understanding of race/gender, which may result in data that is not suitable for the intended purpose. The other is an ethical problem related to the exploitation of the controversial nature of these terms to produce unfair outcomes.

Sacrificing Accuracy for Precision: Why all variables are not created equal (Catherine Greene, LSE)
Criticism of algorithmic decision-making often focusses on fairness and discrimination. While recognising that these are vital, this paper argues that the inaccuracy of the in-puts of such algorithms is also an ethical issue. Wallach writes, "we must treat machine learning for social science very differently from the way we treat machine learning for, say, handwriting recognition or playing chess. We cannot just apply machine learning methods in a black-box fashion, as if computational social science were simply computer science plus social data." (2018, pg. 44). The importance of data accuracy is illustrated by the, now famous, The State of Wisconsin vs Loomis case, in which the Wisconsin Supreme Court rejected a defendant's claim that the use of an algorithm to predict his risk of recidivism violated his due process rights because the defendant had the opportunity to verify the accuracy of the data the system used. The data used to predict his risk of recidivism came primarily from a questionnaire that he had filled out. This paper argues that verifying that a defendant's responses are true, is not the same as verifying that they are accurate for the purpose of algorithmic decision-making. The extent to which offenders belong to criminal groups, and are subject to criminal influences, has been shown to be an important factor in reoffending (Gendreau 1996). The questionnaires that provide the data for recidivism prediction algorithms try to assess this with a few, relatively clear, questions. However, the precise answers to these questions encompass a range of underlying behaviour. Treating these responses as accurate reflections of their susceptibility to reoffend is a mistake. For example, a positive answer to the question: 'Are you a gang member, or associate with gang members?' could mean that a person loosely associates with gang members, or that they are a fully-fledged gang member. This heterogeneity means that the data that algorithms use is not the sort of data that yields reliable predictions. This paper demonstrates the difficulty with ensuring data accuracy using examples from the questionnaires given to offenders, and then presents a framework that data scientists can use to determine when accuracy may be a problem. Even before issues of discrimination arise, it is unethical to treat precise answers to questions as sufficiently accurate to use in algorithmic decision-making. This should be taken this into account when judging algorithmic fairness.

Algorithmic Fairness, Conceptual Engineering, and Controversial Variables (Elizabeth Stewart, University of South Carolina)
Algorithms increasingly play a large role in deciding things about and for humans. In doing so, they must have access to relevant information about those humans lives and identities. However, there has been disagreement over what aspects of human identity are ethical to use in decision making. Multiple studies demonstrate that bearers of historically oppressed identities, including racial and gender identities, receive negatively biased results in algorithmic decisions (Noble, 2018). One of the common objections to the use of these variables is that they can generate

negative feedback loops, wherein the algorithm predicts outcomes while simultaneously causally contributing to those outcomes. In this paper, however, I wish to explore a different way in which the use of these identifiers contributes to unfair outcomes. I argue that variables such as race and gender make poor variables for algorithmic decision making because they have controversial extensions. That is, it is indeterminate what people belong in the extensions of racial and gender terms. While most demographic terms are at least minimally ambiguous, some have more conventionally accepted meanings than others. For example, we know that when we are asked for age, the answer should be given in years, corresponding to the amount of time the earth takes to travel around the sun. Race and gender, however, are two variables that currently lack a stable conventionally accepted meaning. There is currently, both in philosophy and in broader society, an ongoing debate on how best to conceptualize the meanings of these terms. These debates have moved beyond the academic sphere into the public realm, where they have practical, real world consequences. The varying conceptualization of the meanings of race and gender mean that the same individual might, on different accounts of race and gender, be categorized as 'male', 'female', 'white', 'mixed', 'black' or none of the above. The controversial nature of variables such as race and gender leads to several potential problems for algorithmic decision-making. In this talk, I will discuss two problems associated with controversial extensions. The first problem is that the controversial extensions of race and gender can negatively influence data quality. An individual's self-reported gender/race may not correspond to the algorithm developer's understanding of race/gender, which may result in data that is not suitable for the intended purpose. The other is an ethical problem related to the exploitation of the controversial nature of these terms to produce unfair outcomes.

Proxies Aren't Intentional, They're Intentional (Gabbrielle Johnson, Claremont McKenna College)
This talk concerns 'The Proxy Problem': often machine learning programs utilize seemingly innocuous features as proxies for social sensitive attributes, posing various challenges for the creation of ethical algorithms. I argue that to address this problem, we must first settle a prior question of what it means for an algorithm that only has access to seemingly neutral features to be using those features as ‘proxies’ for, and so to be making decisions on the basis of, protected class features. I argue against theories of proxy discrimination in law and political theory that rely on overly intellectual views of the intentions of the agents involved or on overly deflationary views that reduce proxy use to mere statistical correlation. Instead, I adopt an anti-individualist account of representational content to argue for a constitutive account of ‘contentful proxy use’ that draws on resources in philosophy of language and mind. On this view, proxies represent socially sensitive features when and only when they constitutively depend on discriminatory practices against members of marginalized groups.

Fair Decisions, Hard and Soft (Kenneth Silver, Trinity College Dublin & Greg Faletto, University of Southern California)
Those working on algorithmic fairness have offered a number of distinct criteria for fair decision-making. However, it has been recognized that a number of these criteria cannot be mutually

satisfied on any model. A natural response would be to argue for using certain criteria and against others, where our preferred criteria are consistent. Here, philosophers could step in to show how certain criteria do not deliver an appropriate conception of fairness, and how others do. After all, philosophers have already long argued about our conception of fairness. Though reasonable, we worry that it cannot succeed. It may be that these criteria capture distinct senses of fairness, or perhaps they capture other considerations of value. To allow for this, we appeal to a critical distinction in mathematical optimization, and we argue that availing ourselves of it provides for several ways of accommodating apparently inconsistent criteria. In mathematical optimization, practitioners often distinguish between so-called 'hard constraints,' where some mathematical criterion must be satisfied with exact equality, and 'relaxed' or 'soft constraints' where violations of a criterion within a certain margin are tolerated (Boyd & Vandenberghe 2004:Sec.5.1.4). Crucially, the criteria of fairness are inconsistent only if each is treated as hard constraints on the algorithm, and inconsistency can be resolved by relaxing this assumption. So, this paper primarily concerns what considerations suggest criteria for fairness as hard or soft, and how this influences a potential algorithm for fair decision-making. One approach suggests that the criteria for fairness should each be treated as soft constraints. This has philosophical precedence going back to Broome, who treated fairness as a consideration that can be outweighed. If the criteria are not only soft but measurable/estimable, a practitioner can choose appropriate quantitative weights for each measure of fairness (or maximum tolerable violations of each fairness constraint may be chosen) along with a measure of predictive accuracy. Then, a fair model is easily estimated using standard methods. However, there may be reasons why certain fairness criteria should be hard constraints. We show how this comes out of accepting a connection between fairness and respect for certain rights. Rights have been argued to generate exclusionary reasons, and we show how the role played by exclusionary reasons in reasoning maps directly onto the conception of hard constraints. Given this, we suggest a view of algorithmic fairness allowing for mixed constraints, and we offer a roadmap for how to use it.

Should We Believe the Algorithm? Fairness, Epistemic Responsibility and Moral Encroachment (Jannik Zeiser, University of Hannover)

In recent years, the question of algorithmic fairness has rightly received a great deal of attention. I suggest that we need to extend our focus to the issue of responsible belief formation. What we believe, and how we come to believe it, has both epistemic and moral significance. Moral considerations, some concerning fairness, might play a role when it comes to the justification of beliefs informed by algorithms. I will explore this stance using arguments from the debate on moral encroachment (see e.g. Basu 2019; Bolinger 2018, 2020; Fritz 2017). The thesis of moral encroachment holds that the epistemic status of a belief can depend not only on epistemic factors such as quality of evidence or reliability of statistics, but also on moral considerations. For example, in the context of algorithmic processing, imagine an image recognition programme designed to identify different species of plants. When the programme is fed a picture of a flower and identifies it as a daisy, then it seems right to say that I am justified in believing this plant to be a daisy. However, suppose I want to serve a mushroom omelette to a group of friends, and I take it upon myself to find suitable mushrooms. Given that the programme's misidentification of a poisonous mushroom could get my friends injured or killed, I no longer seem to be justified in taking the machine-learning evidence at face value, and relying only on the output of the algorithm to justify my belief in this case. Although the evidence in both scenarios is the same, I only seem to be justified in my belief in the first case. The moral stakes of a belief thus seem to play a role in our standards for justification. Starting from this simple example, I will explore which specific features of algorithmic evidence might contribute to a lack of justifiability in morally loaded situations. This is especially relevant in high- stakes cases, like the COMPAS software which is used to provide risk scores for criminal defendants. Such algorithms have the tendency to unfairly disadvantage some groups, standardize decision procedures, and are unable to treat people as individuals. I will show how these typical features of algorithms tie into arguments from moral encroachment, and how this new perspective may contribute to an improved understanding of fairness demands in AI-based decision-making. This does not rule out algorithms as sources of evidence per se. But it gives responsible epistemic agents reason to seek out evidence beyond that provided by these algorithms.

Attack of the Behaviorist Robots! The Ethics of Reliable Black Box AI (Andrew Knoll, Grand Valley State University)
A relatively under-explored question concerning algorithmic fairness is what I call the Reliable Black Box Problem: Given that an artificial intelligence (AI) can perform a task with as much or greater reliability than a human, is it ever morally permissible (or indeed, obligatory) to deploy it— even when we do not know how it comes to be so reliable? For example, you may wonder whether autonomous weapons, which both select and engage targets on the basis of their programming, ought be deployed on the battlefield (Purves et al. 2015). Or, whether AI programs should be utilized to make parole decisions (Angwin et al. 2016), or direct police to areas where crimes are more likely to be committed (Moravec 2019; Predpol 2020). The Reliable Black Box Problem has us ask how all such questions ought be answered given that the relevant AI is more accurate or reliable than humans, but we do not know how it is so accurate1. The question is both technologically and politically salient. Many of our current AI are reliable black boxes (e.g., He et al. (2016); Mnih et al. (2015); Silver et al. (2018)). Meanwhile, lawmakers have begun drafting policies that regulate such technology. The European Parliament has considered adopting a "right to explanation" that might guarantee individuals information about the process by which an AI had made decisions regarding, e.g., their credit worthiness, job applications, or level of violent threat (Wachter et al., 2017a & 2017b; Hacker et al., 2020). Black box AI would by definition not allow for such explanation. This presentation aims to clarify just what it means for an AI to be "reliable" and what it means for it to be a "black box" so that we can draw better informed ethical conclusions about its deployment. I'll clarify the notions of reliability and black box in terms of functions in extension versus functions in intension, following Church (1941). An AI is reliable only relative to a specified function in extension. Specifically, an AI is reliable relative to a function to the extent that its behavior instantiates that function. An AI is a black box to the extent that we are ignorant of the algorithmic function in intension that implements the function in extension. These conceptual clarifications allow me to argue that, all things being equal, we are morally obligated to use an AI that is more reliable but less algorithmically transparent rather than one that is less reliable but more transparent. We may well morally err in choosing a

function in extension for an AI to carry out— but assuming that our choice of function is just, then the more reliably an AI instantiates it, the better. That's true even if we remain ignorant of the algorithm— the function in intension— that implements it. Of course, against my claims, some argue that a person is wronged if they cannot receive an explanation as to why they were denied a loan, injured in war, or denied parole. But, I argue that the very reliability of an AI ought be explanation enough to satisfy such demands.