Dialectical AI: Overcoming the Eugenicist History of Probability

by Ken Archer

1. Introduction

This paper argues that the ethical concerns raised by AI are, at a fundamental level, continuous with those throughout the history of statistics, and chiefly concern the role of probabilistic models within the scientific advance of social practices. The role of models in human practices has been in question since the birth of classical probability. Are models a formalization of limited human reasoning that progressively free themselves of human bias and inconsistency until they function relatively autonomously and prescriptively? Or are models embedded within a social dialectic through which statisticians and human domain experts iteratively, collaboratively advance intelligence within a practice, making a practice more scientific?

Each of these two roles of probabilistic models carries with it assumptions about the nature of human reasoning, the problem that is solved by probability and the resulting ethical vision of how probability solves this problem. In the former case, which we'll call formalist probability,[1] human reasoning reduces uncertainty - understood as quantifiable doubt - through induction, and probabilistic models solve for the imperfections of bias, inconsistency and limited memory that plague human induction, through mathematical formalization. Humans err by inducing patterns where there is really random variation, and mathematical formalization of random variation enables science to distinguish true correlations from random chance. The ethical vision that animates formalist probability is thus circumscribed to a freedom from human limitations through formalization, while remaining agnostic to how formal models are employed in human practices.

In the latter case, which we'll call dialectical probability, human reasoning is broader, and characterized by two types of uncertainty – the reduction of quantifiable doubt through induction as understood by formalist probability, but more fundamentally the reduction of ambiguity through clarification of how a domain should be conceptualized in the first place. Ambiguity and doubt are thus two axes of uncertainty according to dialectical probability, whereas formalist probability reduces all uncertainty to quantifiable doubt.[2] Human error results primarily from ambiguity. Random variation, of the type we associate with lotteries, may or may not exist in the world – on this dialectical probability is silent, as the purpose of probabilistic models is to stimulate clearer thinking from domain experts on how to specify, classify and make explicit the causal relations within their domain. They do this by suggesting ways to resolve ambiguity and helping verify attempts to do so. This in turn leads to more clearly specified models, and clearer specification of the data gathering process that is inseparable from model development, in a virtuous dialectic through which scientific knowledge is developed within a domain or practice.

[1] Others have diagnosed the underlying ethical problem of AI as algorithmic formalism (Green 2019) and the formalism trap (Selbst 2019). Ben Green and Salome Viljoen, "Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought", ACM, *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, doi: 10.1145/3351095.3372840. A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi., "Fairness and abstraction in sociotechnical systems", ACM, *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, doi: 10.1145/3287560.3287598.

[2] Knight and Keynes in the early 20th century were notable for deepening our understanding of uncertainty as both doubt, or risk, and ambiguity. Frank Knight, *Risk, Uncertainty and Profit*. (1921, Cambridge: The Riverside Press). John Maynard Keynes, *A Treatise on Probability*. (1921, London: Macmillan).

The ethical impetus towards dialectical probability, in contrast to formalist probability, is not autonomy from human limitations, but the development of practical knowledge itself, making practical knowledge more scientific. Model autonomy is not a value, and so success is not defined in terms of autonomy from human participation in a practice, but in terms of the advance of the practice itself. While the use of models for autonomous decision making occurs within a more knowledgeable practice, so does the clarification of practical knowledge and judgment by practitioners.

The tension between these two understandings of probability runs through the history of probability and statistics and underlies ethical debates within this history, from eugenics to the crisis of replicability in science to the harms of AI bias. This is because those engaged in these debates are essentially talking past each other, not only conferring different meanings to central concepts like uncertainty and probability, but also conferring competing values to these differing meanings.

Notice, for example, that both formalist and dialectical probability lay claim to being objective. However, their notions of objectivity are clearly different. Whereas the dialectical notion of objectivity – truth-to-nature – seeks to resolve the problem of ambiguity, the formalist notion of objectivity – mechanical objectivity – sees the problem more broadly as subjectivity itself and solves this problem through mechanistic approaches to knowledge creation that eliminate the self entirely.[3] Each notion of objectivity has a corollary subjectivity that constitutes its latent ethical content – the scientific advance of practices, on the one hand, and freedom from human limitations, on the other.

Rather than appealing to "ethics" as an external concern that "AI" must acknowledge, then, this paper acknowledges the latent ethical content of what it means to be a good statistician or engineer within AI, and simply calls for AI workers to do better AI – dialectical AI – while appealing to the long history of statisticians and probabilists who make the same appeal.

Advocates for AI ethics tend to teach a set of exogenous principles – fairness, accountability, trust, privacy - that statisticians and engineers are expected to apply, on the assumption that predictive model accuracy is in tension with ethical constraints. The implication of such advocacy is that, whereas AI systems may not be morally neutral, those who build them *are* applying instrumental, calculative reasoning - formalist probability - to build towards a design and must be taught the ethical ramifications of their finished designs. AI researchers have responded in turn with a mathematicization of ethical concepts, like fairness, as formal constraints on models optimized for accuracy.

This formalist framing of AI unwittingly adopts a self-understanding of statistics work that AI ethics advocacy should in fact resist. By buying into formalist assumptions around probability, AI ethics perpetuates the self-understanding of technical work within AI as an applied science which allows little room for agency, such that builders of AI systems have agency primarily in the decision *whether* to build something. Once that decision is made,

[3] These notions of objectivity are given historical treatment in Lorraine Dalston and Peter Galison's *Objectivity*, 2007, Zone Books, Brooklyn, NY.

however, the agency of the builder is narrowed significantly to the instrumental application of statistics and computer science.

Dialectical probability adopts a wider frame of reference, within which there is no trade-off between accuracy and ethics. The relevant frame of reference broadens from an algorithmic frame to a sociotechnical frame, of which the algorithm is considered a subsystem. The standard of success for a model isn't autonomy from the limitations of human cognition, but more broadly the advance of rational knowledge, in service of which models are a useful tool.

As the ethical concerns around AI have mounted, not all AI researchers within AI ethics have responded with mathematical formalizations of ethical concepts. An emerging critique from within AI diagnoses the underlying ethical problem of AI as algorithmic formalism[4] and the formalism trap[5]. While these critiques likewise advocate for widening the frame of reference of AI to sociotechnical systems, they do so by appealing to fields outside of statistics and computer science (specifically, law and STS) that, while illuminating, fail to appeal to resources from within AI of what it means to do AI well.

By appealing to what it means to do AI well, this paper draws on the long tradition of virtue ethics, and shares with MacIntyre an appeal to the internal goods of practices as the basis for ethics. Ethical appeals gain more traction when they are made in terms of, and not in opposition to, our existing inclinations for engaging in a practice.

A common reason given for the need to incorporate ethical concerns into an otherwise instrumental technology such as AI is the claim of unintended consequences. Engineers can't anticipate the ways that technologies will be used by others.[6] This line of argument, too, buys into a formalistic understanding of how technology is built that is in fact contested within statistics and within other technical practices. Engineers and statisticians don't need to yield their authority within their field of expertise to ethicists; they just need to be better engineers and statisticians.

This paper thus argues that technical rationality is never merely calculative and is always already embedded in self-interpretive practices with competing goods, such that technical decisions are more authentically understood as hermeneutic self-interpretations of practical predicaments. The critical role of technology ethics, it seems to me, is to unpack and account for the full ethical content of the decisions faced by the statistician and engineer, including the competing internal moral visions of the technical practitioner at play in these decisions. Such an account would restore the sense of ethical agency that the formalist notion of technology as applied science conceals from engineers and statisticians.[7]

[4] Green 2020

[5] Selbst 2019

[6] This was the central argument of Hans Jonas in *The Imperative of Responsibility*.

[7] This paper is thus situated within the concern for subjectivity, and the freedom and agency of the subject, that originates in Rousseau and Kant and continues through Romantic and then hermeneutic thought, most recently found in Gadamer, Ricoeur and Taylor. This is in contrast to postphenomenological approaches to technology ethics that seek to break from what is described as transcendental understandings of technology in favor of empirical research into the ways that artifacts condition our lives. What is critiqued as transcendental understandings of technology, primarily the later Heidegger, is actually part of this longer

Calls for ethical guidance of technology from the outside – via ethics training, via roles assigned to humanities professionals and via regulatory constraints – carry a tinge of elitism and reinforce the narrow agency afforded engineers. By avoiding engagement with specific technical decisions, ethics as an external guide to technical reasoning even enables ethics-washing of such decisions. Nothing less than an appeal to competing internal goods within technical reasoning itself, and a concomitant reframing of technical knowledge and training, will change the current trajectory of technology and of modernity. This may be a tall order, to disclose a moral compass from within the most technical areas of practice, but there is no way around it.

2. Formalist and Dialectical AI

The mounting ethical concerns raised about AI are more properly understood as an opportunity to widen our understanding of AI, from the formalist AI that provoked the unintended consequences to which the AI ethics community is responding, to a dialectical AI conceived as a process of which an algorithm is one part. Whether this widening occurs depends on whether the response to these concerns is to coopt them within formalist AI, or to do better AI.

Two demands in particular – fairness and explainability – are scenes of this tension. The response of formalist AI to these demands is to characterize them as independent of the sole concern of AI for accuracy, and thus as requiring a trade-off – between accuracy and fairness, or between accuracy and explainability. When AI is viewed more broadly as a sociotechnical process, rather than as an algorithm, then demands for fairness and explainability are no longer framed as separate from the concern for accuracy; these demands actually point to the essential role of social context in creating accurate algorithms.

Fairness

The well-known debate about the use of AI in sentencing decisions within the criminal justice system illustrates the harm that results from formalist approaches to AI, including the formalist incorporation of ethical concepts themselves. Algorithms for recidivism prediction carry a lot of promise. By making the practice of bail setting and judicial sentencing more evidence-based, we could reduce the amount of crime and/or reduce the population of incarcerated people.[8]

However, this promise carries with it the risk that recidivism prediction algorithms will simply encode existing biases in arrest patterns. In 2016, investigative journalists for ProPublica found evidence that appeared to suggest exactly that. According to ProPublica's investigation, the COMPAS algorithm for recidivism prediction produces higher false positive rates for black defendants than for white defendants.

The makers of COMPAS replied that this was the wrong metric for measuring fairness. The algorithm generated risk scores that were calibrated. That is, for persons with the same risk

tradition of concern for the freedom of the subject against the heteronomous demands of empiricist and rationalist epistemologies.

[8] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions (No. w23180). National Bureau of Economic Research.

score, blacks and whites had the same propensity for recidivism and knowing one's race would add no more information.

Both sides, it should be noted, appealed to formal metrics of fairness. In recent years, ML fairness researchers have designed several mathematical formalizations of fairness. The most common are parity of false positives and false negatives – known as equal odds – and calibration.[9] A substantial portion of ML fairness research across top universities and tech companies has focused on developing these formalizations of fairness: accounting for fairness within a large number of intersectional groups, accounting for fairness when sensitive group attribute data is not available, and so on.

Fairness research is thus happening within the *algorithmic frame* of formalist AI. Within this frame, the relevant features are the output, the training data, and the relationship between them. Success is defined narrowly in terms of accuracy of the output in relation to the training data, and generalizability to new data from the same distribution. This artificial context comes with no concepts for expressing ethical notions such as fairness, and so fairness can only be expressed as a regulatory constraint on such an activity or an algorithmic constraint.

When faced with the question of regulatory constraints, AI researchers frequently appeal to the ethical vision of formalist AI as overcoming human limitations, an appeal that always begins with reducing human reasoning to algorithmic induction.

> *First, all decision-making – including that carried out by human beings – is ultimately algorithmic. The difference is that human decision-making is based on logic or behaviors that we struggle to precisely enunciate. If we humans had the ability to describe our own decision-making processes precisely enough, then we could in fact represent them as computer algorithms. So the choice is not whether to avoid using algorithms or not, but whether or not we should use precisely specified algorithms. All things being equal, we should prefer being precise about what we are doing.*[10]

As is shown below, the premise that all decision-making, human or machine, is algorithmic, is precisely what is being challenged by demands for fairness and explainability, and in fact has been challenged through the history of probability and statistics.

Another reason AI researchers oppose regulatory constraints is because, having reduced human reasoning to algorithmic induction, it makes sense to believe we can similarly reduce notions of fairness, by mathematically formalizing fairness as a constraint on the algorithmic relationship between training data and the output.

As has been shown by researchers working in decision theory, however, these formalizations are poor measures for detecting discriminatory algorithms and can negatively

[9] Hardt, M., Price, E., and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems.* 3315-3323. Aaron Roth and Michael Kearns, *The Ethical Algorithm*. 2019.

[10] *The Ethical Algorithm*, p. 191

impact marginalized groups when used to design fair algorithms.[11] Decision theory broadens the frame of reference within which algorithms are designed and evaluated from an algorithmic frame to a *sociotechnical frame* that encompasses the real-world harms and benefits of decisions made with an algorithm.

The harm of a false positive in recidivism prediction is unnecessary incarceration, while the harm of a false negative is avoidable crime. Once the costs and benefits of incarceration are included within the frame of reference, then algorithm designers must consider the optimal threshold that balances these costs and benefits. Perhaps the benefit of preventing a crime is twice the cost of unnecessary incarceration. Perhaps the ratio is different for different types of crime, or for persons who are primary caregivers for children.

Once a threshold rule is identified that balances the costs and benefits of incarceration, one would expect a fair algorithm to use the same threshold regardless of protected group status such as ethnicity. The costs of incarceration, and of wrongful incarceration, are presumably the same regardless of ethnicity. This is not to argue that recidivism prediction algorithms are not biased, just that reducing fairness to the narrow frame of an algorithmic formalization is a poor way to detect bias.

In fact, applying the same utility threshold fairly across protected groups generally results in different false positive and false negative error rates across groups. To see why, we have to consider risk distributions, which make clear the need to look at cases close to the threshold or margin - individual cases - in order to improve our understanding and measures of fairness.

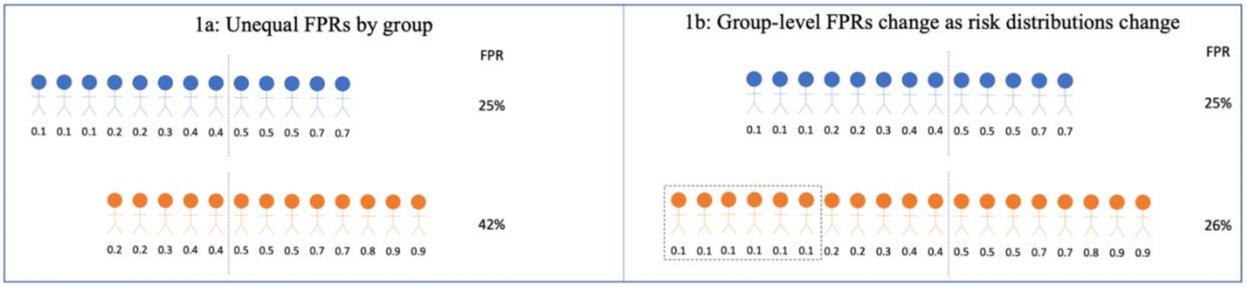


Figure 1: The Problem with False Positive Rates. Courtesy of Sam Corbett-Davies, et al, from Making Fair Decisions with Algorithms.[12]

Figure 1 illustrates the problem with using equal false positive rates – sometimes called equal opportunity - as a measure of fairness. The false positive rate is the number of false positives – in this case the number of unnecessarily incarcerated defendants – divided by the total number of negatives - defendants who will appear for their hearing without committing crimes. (FPR = False Positives/(False Positives + True Negatives).) A risk assessment algorithm has assigned recidivism risk scores to defendants in two groups. In Figure 1a, the defendants in the orange group have a higher mean risk than those in the blue group, and a

[11] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 535-545; Sam Corbett-Davies and Sharad Goel. 2018. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.*

[12] Sam Corbett-Davies et al. https://samcorbettdavies.files.wordpress.com/2017/11/making-fair-decisions-with-algorithms.pdf

threshold of 50% is being used for bail decisions. As a result, the false positive rates are different between the two groups. This is a simplified form of how recidivism prediction algorithms like COMPAS are operating.

Let's say that police start enforcing minor crimes disproportionately impacting the orange group, adding several low-risk persons. As is seen in Figure 1b, this makes the false positive rates equal between the groups. But nothing has actually changed, except for the risk distribution within each group.

The distribution of risk scores by race observed by ProPublica journalists, in Figure 2, displays the same dynamic.

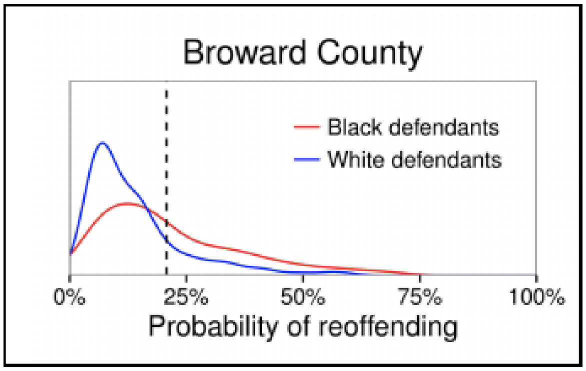


Figure 2: Distribution of Risk Scores by Race from COMPAS Recidivism Algorithm

The means for black and white defendants differ, which by definition results in differing risk distributions and different false positive rates when a single threshold is applied fairly across all groups.

The focus on the decision threshold that optimally balances the costs and benefits of false positives, true positives, false negatives and true negatives broadens the frame of reference in two ways.

First, rather than focus on formalizations that reduce fairness to the algorithmic frame, we must consult domain experts, judges in this case, for their definition of fairness in terms of these trade-offs, and their determination of whether the trade-off is the same for everyone across protected groups.

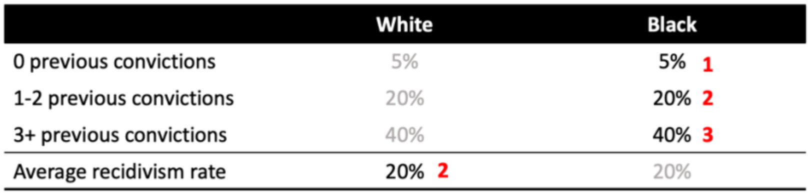
Second, what we learn from domain experts is that considerations of fairness do not arise in many cases – in cases where the probability of reoffending is clearly high or low, the decision to be made is clear regardless of group status - but one that comes to the fore in individual cases on the margin. So, like with algorithmic fairness, fairness in human decisions is relevant for cases close to the margin. This makes intuitive sense but is overlooked by formalization of fairness that abstract across all individuals within a protected group.

Furthermore, we learn from domain experts that they have more information than the algorithm in cases close to the margin – e.g. judges talk to individual defendants. We might learn that the cost-benefit trade-off is different when the defendant is the sole provider for a child, which is more likely to be the case with female defendants.

In other words, we learn that fairness, like many properties of real-world domains, is inherently ambiguous. The role of AI in judicial sentencing is thus not just to correct for human limitations of bias and limited memory, but also to further clarify what fairness means, a process that likely never will reach closure. However, given that this dialectical exchange between the AI designer and the judge is one that occurs for cases on the margin, this immediately suggests an efficient process for making the practice of judicial sentencing more rational. First, for cases far from the margin, permit near-automated decision making, and review all cases where the judicial sentence overrides the algorithmic recommendation. Second, for cases close to the margin, require human decision making based on a score recommendation, as well as a review between the AI designer and judges of the cost-benefit considerations brought to bear for a random sample of cases near the margin.

False positives and false negatives aren't the only reductive formalizations of fairness used in fairness AI research. Calibration, the metric to which the maker of COMPAS appealed, likewise fails to capture the full meaning of fairness.

A model is calibrated if defendants with the same risk score reoffend at the same rate across groups. While any fair model would certainly be calibrated, this is a fairly low bar. Consider the example in Figure 3, in which white and black defendants reoffend at the same rates conditional upon previous conditions. However, while the risk scores for black defendants factor in previous convictions, assigning black defendants risk scores from 1-3, the risk scores for white defendants do not, and all white defendants are assigned a risk score of 2.



	White	Black
0 previous convictions	5%	5% 1
1-2 previous convictions	20%	20% 2
3+ previous convictions	40%	40% 3
Average recidivism rate	20% 2	20%

Figure 3: Calibration is a weak guarantee of fairness. Courtesy of Sam Corbett-Davies, et al from Making Fair Decisions with Algorithms.

This model is calibrated, as blacks and white with risk scores of 2 reoffend at equal rates. However, when using a decision threshold of 25%, no white defendants are detained. By failing to consider factors that would discriminate white defendants from each other, this model is unfair to black defendants. This is essentially the phenomenon of redlining, in which, historically, black mortgage applicants were judged creditworthy based only on their neighborhood, and individually discriminating factors like income were ignored.

The solution to redlining is to inspect the model to ensure all features relevant to each group are included. This sometimes requires the inclusion of protected group status, like race, in the model, as features are often differentially predictive within each group.

As with the determination of cost-benefit thresholds above, the inspection of a model to ensure features that differentiate members within each group are included relies upon consultation from domain experts. There is no simple fairness metric, whether calibration or equal error rates, against which an algorithm can be optimized to ensure fairness. In fact, as we have seen here, such reductive formulations of fairness serve to conceal bias.

Two other sources of bias cannot be detected through algorithmic checks and require broadening the frame of reference to the sociotechnical frame. *Label bias* occurs when the outcomes – such as arrest rates – in training data are inconsistently applied between groups and thus unreliable. For example, if one neighborhood is more heavily policed, then crimes are more likely to be detected. But recidivism algorithms aim to predict criminal activity, not simply arrests.

Sample bias, likewise, cannot be detected with algorithmic checks. Sample bias occurs when the population in which an algorithm is used to score and make decisions differs in relevant ways from the population from which the training data was pulled. For example, applying a recidivism algorithm trained in one jurisdiction to another jurisdiction, or implementing the algorithm as part of a bail reform campaign within a jurisdiction, results in non-representative data used to train an algorithm.

Sample bias is alleviated in two ways. First, domain experts pay careful attention to relevant changes in the training and target populations of an algorithm. Second, algorithms are inspected by domain experts to ensure features are generalizable from one population to another, and don't simply reflect patterns specific to the population from which the training data was drawn. Features that transfer more easily across populations are those with a more causal relationship to the outcome, rather than those that happen to correlate with the outcome in a specific population.

Inspection of algorithms by domain experts, to address redlining and sample bias, is thus essential for algorithmic fairness. This requires some degree of model explainability, the topic to which we turn next.

Explainability

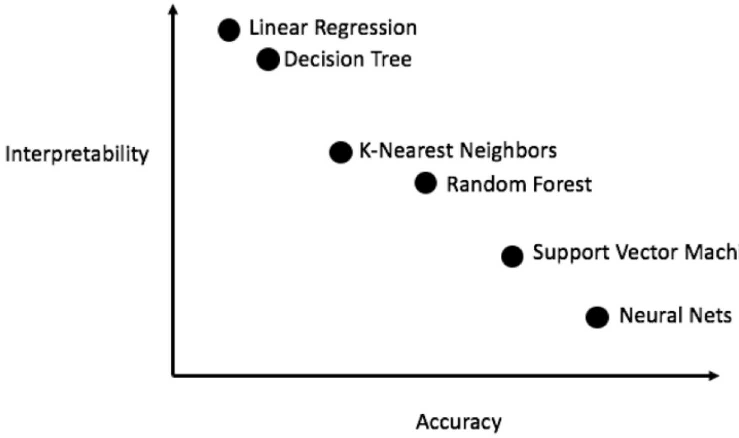
The competing approaches to AI are seen in another domain - hospital admission algorithms. In the 1990s, a national effort was undertaken to build algorithms to predict which pneumonia patients should be admitted to hospitals and which treated as outpatients. Initial findings indicated neural nets were far more accurate than previous statistical methods. Some researchers were concerned about the use of such models and looked deeper to see how they were working.

It turned out that the neural net had inferred that pneumonia patients with asthma have a lower risk of dying and shouldn't be admitted to the hospital. Obviously, this is counterintuitive to anyone with domain expertise. But it reflected a real pattern in the training data—asthma patients with pneumonia usually were admitted not only to the hospital but directly to the ICU, treated aggressively, and survived.[13]

Only by making the model explainable was a crucial problem discovered and avoided and, notably, the model improved.

Many introductions to machine learning algorithms present diagrams such as the one below that claim to illustrate a trade-off between accuracy and interpretability of models.

[13] "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," Caruana, et al. (2015), http://dx.doi.org/10.1145/2783258.2788613.



This alleged trade-off, however, is premised on a formalist notion of accuracy that prescinds from any real-world context in which accuracy actually matters. When AI is viewed as a dialectical process, through which models increase in accuracy over time, then making models transparent to domain experts for debugging becomes critical to improving accuracy. Within formalist AI, however, the accuracy of a model is limited to the success of a model in predicting a target variable, using a loss function, within a test data set, using an accuracy metric. This is essentially a “lab” definition of accuracy that produces numerous inaccuracies in a real-work context.

Limitations of Training Data

First, accuracy is limited to the performance of a model in a test data set, not in the real world. This was mentioned in the discussion of fairness as sample bias. Model performance generally falls off to varying degrees when deployed to the real world, as a test data set is simply a subset of a training data set and thus shares in the spurious associations in the training data set that may not appear in the real world indefinitely. While a test data set is fixed, real-world data is dynamic and non-stationary, and thus limitations in training data inherently limit the real-world robustness of models into the future.

One common response to the concern about spurious associations is that with randomized experimental data models correct for spurious connections. However, randomized experiments are limited in their ability to surface the myriad underlying functional relationships that shape any domain. Furthermore, when applying models to decisions with significant impact on people’s lives, it is rarely even ethical to randomize the treatment given to people (whether that be medical care, court sentencing, or other decisions with sizeable impact on peoples’ lives) and thus models must be trained with the observational data that is available.

Limitations of Loss Functions

Second, accuracy is based on performance in predicting a single target variable. This creates the potential for mismatched objectives, multi-objective trade-offs and spurious associations of features with that target variable that would put vulnerable or protected populations at risk if used in a decision model. All these model misspecifications increase accuracy, using this narrow technical definition of model accuracy “in a lab”, while creating unintended consequences in the real-world decisions made based on models.

Accuracy, then, as presented in the accuracy-interpretability trade-off, has a narrow, technical meaning within the field of machine learning that in no way precludes a highly accurate model functioning far differently than intended when deployed for real-world decision making. In the case of the pneumonia study, a spurious association between asthma and pneumonia risk would have put the lives of asthma patients at risk.[14]

What does interpretability look like?

Spurious model misspecifications can be detected only when models are explainable to domain experts.[15] The purpose of interpretability is thus improved model performance through iterative model debugging and model checking by domain experts.[16]

The purpose of interpretability is frequently assumed to be trust, and while this is true as far as it goes, it is limited as it assumes an interpretable model always deserves to be trusted. As was seen in the pneumonia example, trust in a model is the result of a process of iterative checking and debugging of an interpretable model, not interpretability itself.

Interpretability can mean different things to different people. Is a model interpretable if you can open it up and look at its form and its parameters? If the purpose of interpretability is to enable a process of iterative model checking and debugging, then what we mean when we speak of interpretability becomes clearer, as this purpose requires that humans understand the cause of a decision made by an AI system.[17]

Truly explainable models thus facilitate an iterative relationship between domain experts and AI models that improves models and improves domain expertise.

3. The Historical Roots of Mechanical Objectivity in AI

[14] Different decision domains vary in the limitations of statistical accuracy when compared with real-world, substantive accuracy. Decision domains in which deep learning neural networks have been most successful – image classification, text classification – don't have the limitations on training data and loss functions to the same extent as decision domains in which significant effects on peoples' lives result from the decision. In the latter domains, the potential gap between "lab accuracy" and real-world substantive accuracy is greater.

[15] "The first step towards improving an AI system is to understand it's weaknesses. Obviously, it's more difficult to perform such weakness analysis on black box models than on models which are interpretable." "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", Samek, Wiegand and Muller, https://arxiv.org/abs/1708.08296, August 28, 2017

[16] Hence the complimentary purposes of Explainable AI in *ibid:* "Verification of the system", "Improvement of the system", and "Learning from the system".

[17] Explainable AI is generally divided into two approaches – intrinsically explainable models and post-hoc, model-agnostic models. When AI researchers appeal to a trade-off between interpretability and accuracy, they generally have in mind the interpretability of the model itself. So, linear regression and decision trees are generally more interpretable, though less accurate, than random forests and neural networks. However, growing research into post-hoc models requires no trade-off at all between accuracy and interpretability. Post-hoc methods train explainable models to mimic non-explainable models, thus exposing the associations being made by the model for model debugging and improvement. These post-hoc methods have the additional benefit of being model-agnostic, which means they place no constraints on the opacity of the primary predictive model.

This tension between two understanding of what it means for probabilistic models to make human practices more objective and scientific, and the ethical lapses that have ensued when the formalist understanding becomes predominant, is not new to AI. It underlies the crisis of replicability in science and the late 19th and early 20th century eugenics movement.

It's common to treat these ethical lapses as personal failings of specific scientists that don't implicate statistics in any way. But this characterization of history presumes and advances the formalist notion of AI, according to which the model is completely independent of its real-world use, that is responsible for these ethical lapses in the first place. Until statistics owns these ethical crises as the result of a reductively formalist understanding of probability, they will recur in different forms.

Classical Probability: Grounding Social Progress in Universal Reasonableness

Probability in its classical incarnation was essentially a concept in tension between a broader, descriptive probability of the early probabilists that encompassed both ambiguity and doubt, and a progressively narrower, prescriptive probability that reduced all error to random variation.

The critical backdrop to the 17th century emergence of classical probability, according to the pioneering work of Lorraine Dalston, was the conviction that civic and commercial order comes not from a transcendent order reflected at different levels of being and enforced by the Church, but from the mutual recognition of men as reasonable through contracts.[18] The ambition of the early mathematicians of probability was not to formalize models free of human bias, but rather to uncover and describe in formal terms the unconscious intuitions of reasonable men. Such a recognition of the universal reasonableness of men, it was believed, would ensure a new, secure basis for social order free of the conflict and skepticism that defined the 17th century.[19]

This discovery was ultimately made, as Herbert Weisberg argues, by taking one such aleatory practice – the lottery – as a metaphor for other such practices[20]. Historians have long struggled to explain why the early probabilists were so consumed with problems of gambling. Weisberg demonstrates that it was the metaphorical role of the lottery that enabled classical probabilists to account for the reasonableness of men, all of whom approach an uncertain situation *as if* it were a lottery, in terms other than the role of fortune in the natural order. Specifically, according to classical probabilists beginning with Jacob Bernoulli, reasonableness reckons with chance by abstracting from our ambiguous intuitions a set of specific causes, which

[18] Lorraine Dalston, *Classical Probability in the Enlightenment*, (Princeton University Press, 1988). Like most other advancements in the Scientific Revolutions of the 16th and 17th centuries in the mixed mathematical disciplines of mechanics, astronomy and optics, probability was not essentially a discovery based on new evidence, but a new account of existing evidence, in this case of chance events, that no longer appealed to an overarching natural order. As such, it was the ongoing hold on the early modern imagination of Fortuna as part of the natural order, even as late as Pascal in the 17th century, that held back the discovery of classical probability.

[19] The allure of method that captured Descartes and Bacon inspired Leibniz that controversy could be resolved through reasonable calculus "Let us calculate, Sir; and thus by taking to pen and ink, we should settle the question".

[20] Herbert I. Weisberg, *Willful Ignorance*, (John Wiley, 2014)

are assumed to function *as if* they were a lottery yielding odds for and against an outcome, odds which then dialectically turns back on and clarifies our intuition and judgment.

The Emergence of the Ideal of Objective Validity and Freedom from Bias

A tension within this classical account of probability has animated debates about the proper role of probability from its beginnings to the present-day. Does probability describe how people reason in the face of uncertainty, or prescribe how they should reason? When descriptive accounts of decision making clashed with the actual decisions made by most people, probabilists cast uncertainty no longer as sensible reasoning that is refined by probabilistic abstraction, as if a situation could be conceived as a metaphorical lottery, but instead reduced uncertainty to mathematical probability. Uncertainty no longer encompassed ambiguity into the causal setup of a situation but was instead simply identified with random variation – a random lottery - as measured by mathematical probability.[21]

Thus began the emergence of the moral ideal animating statistics that today is paramount, the search for a universally valid position that frees people of bias and prejudice.[22] Whereas the first era of classical probability sought to inform judgments with models of formalized good sense, what is definitive of the 2nd era of probability, extending from roughly 1840 to the present-day, is the rejection of judgment and subjectivity itself.[23] The moral vision of enlightened judgments by the many was replaced by a moral vision of universally valid knowledge that eliminated subjective prejudice through mechanistic objectivity. As explained below, this formalist probability would lead to multiple ethical crises, from eugenics to the replicability crisis, and would inevitably cause more harm when applied to whole new practices in AI.

The mean became the chief object of statistical analysis. Rather than the rationality of the few as the model, statisticians turned to the irrationality of the many and made the average person the objective basis for statistics. The term 'statistics' itself arose in the late 1700s as the

[21] From the beginning of classical accounts of probabilistic reasoning, there was a tension between descriptive accounts of reasonableness, on the one hand, and resistance to such accounts by social actors, on the other. Resistance was encountered from both those whose actions defied the accounts of the probabilists, and those who objected to the probabilist accounts of their good judgment. This was true of sellers of insurance and annuities as of gamblers, both of whom rejected models custom-made based on observation of their practices. Practitioners in areas involving risk and judgment viewed their expertise in terms of seasoned judgment of the individual case – not in terms of rules that applied en masse. This tension threatened the moral assumption of reasonableness as the basis of social order, a threat that led to a more concretized understanding of thinking-as-a-lottery that took on a more prescriptive role. This reification of the metaphorical lottery is the basis of the ongoing debate as to its location – whether probability exists in objective reality or subjectively in our minds.

[22] In the middle decades of the 18th century, a hope for the scourge of smallpox that killed 1/7th of Europe came in the form of an inoculation practiced in Turkey that appeared effective, but in the short-term was found to kill about 1 in 200 of its recipients. When Daniel Bernoulli submitted a paper in 1760 to the Paris Academy of Sciences using probability to combine the short-term and long-term risks of inoculation and of no inoculation, to his disappointment many people continued to choose the long-term risk.

[23] For the history of modern statistics from the 19th century see *The Taming of Chance*, Ian Hacking (Cambridge University Press, 2008) and Chapters 2-3 of *The Taming of Chance.*

social mathematics of states, as the focus of analysis shifted from understanding individual decisions to understanding society.[24]

This shift was reflected in Poisson's revision of Bernoulli's law of large numbers to account not for a single underlying causal probability, but for fluctuating underlying probabilities, whose effects were found to converge to a mean more quickly than those of a single probability. Unlike classical probability which assumed a mechanistic causal order that was probabilistically understood by people of good sense, 19th century probability saw lawlike regularities as arising from disorder and a myriad of various causes. The regularities in births, marriages, crimes and suicides on display in the "avalanche of numbers" produced by new statistical offices across 19th century Europe pointed to a deep social order controlling what had been assumed to be random or anti-social activities. For advocates of statistics such as Quetelet, this deeper social reality, attested by the stability of mean values, is more real than the individuals counted, and he called for responsible government intrusion to cure anti-social maladies like crime and suicide now known to be under social control.

Quetelet was an astronomer and saw himself as bringing the mathematical study of physics to "social physics" which had heretofore been the backwater of a subjective calculus of reasonableness. He had learned the formula of Gauss governing the distribution of errors from true values in astronomical observations and applied it to social statistics. Quetelet and Francis Galton were deeply influenced by the application of the Gaussian distribution, known later as the normal distribution or bell curve, to society, and became committed to eugenics. Whereas Quetelet saw the average man as the type of a nation in comparison to which individuals were flawed, Galton saw the upper ends of the curve as the variation that was the hereditary source of genius. Challenged to explain why the offspring of exceptional people would ultimately revert back to the mean of their ancestors, Galton developed the methods of correlation and regression to attribute the variation of offspring in part to one's parents and in part to variation in the offspring.

It's critical to clarify the specific relationship between eugenics and statistics. At one extreme, one could argue that eugenics was a personal failing of great statisticians, but the personal failure doesn't implicate statistics. At the other extreme, one could argue that statistics is fatally flawed, as it was designed for eugenic aims and is only useful for similar projects of discrimination and control. The argument here takes a third route – statistics is beset by an internal tension, and it is the dominance of the formalist approach to probability that accounts for eugenics, the crisis of replicability in the sciences and the harms of AI.

Specifically, in the case of eugenics, by making random variation a part of objective reality, and no longer a convenient metaphor to describe and validate what others took to be the causal setup of their domains, statisticians could dispense with domain "experts" now dismissed as limited by subjective biases and reduce science to a selective abstraction of correlates with mean effects. The distribution of errors around such mean effects were no longer the result of ambiguity, which we would need domain experts to resolve, but random variation.

[24] *Seeing Like a State,* James C. Scott (Yale U Press, 1999)

This reductive formalization of error allows the statistician to play a two-fold role: while claiming to be free from bias, they are now the ones selecting correlates whose main effects are privileged in scientific analysis, analysis that is shielded from criticism by claims to mechanistic objectivity. The two-fold role is what allows the biases of statisticians to influence scientific analysis, while also preventing them from seeing such biases under the guise of mathematical objectivity.

From a broader understanding of probability as a dialectical process, there is no reason why main effects should be primary at all in the causal setup of a domain. This is known as the main effects fallacy, in which we first determine main effects, and then add on interaction effects of those main factors.[25] However, in social domains with complex interlocking causes we would expect interaction effects to be primary.

The seemingly objective reality of main effects with random errors convinced Galton that his mathematics of correlation and regression could explain a range of phenomena beyond biological inheritance. He called for an independent, objective mathematical statistics that could be applied to any field. Karl Pearson and R.A. Fisher, also concerned with eugenics, shared Galton's vision and together developed in the early decades of the 20th century much of the mathematical statistics that we know today. The mathematization of statistics over the first half of the 20th century, which dominates present-day statistics, thus emerged directly from the deeper reality of main effects and random error that eugenicists saw as underlying phenomenal appearances.

The main effects fallacy, premised on error being due to mathematically measurable random variation and not ambiguity, made statistics the vehicle for advancing eugenics. Pearson created the Chi Square significance test to measure the distance between two distributions, such as that between different racial groups, as conclusive evidence that the chosen racial correlates are the causal setup that explains variation in intelligence.[26]

The result was a field with an identity crisis. Statistics was a part of mathematics that was prior to, and thus superior to, any particular application, while at the same time its *raison d'etre* was its value to society and to decision making, and any such application of statistical theory required multiple judgments informed by domain expertise. When such judgments are not transparent, they are the means through which bias – whether that of the eugenicist, the scientist publishing non-replicable research, or the AI worker inflicting harm on marginalized communities – is concealed behind a façade of mechanical objectivity.

This identity crisis is on full display in the statistics of hypothesis testing. Fisher developed Pearson's significance testing in order to provide a scientific method for the experimental evaluation of all hypotheses. Fisher's method of significance testing evaluates the null hypothesis that two distributions of observations – one from a randomized sample that receives no treatment and another from a randomized sample the receives an experimental

[25] Weisberg, pp. 316-317

[26] Aubrey Clayton, "How Eugenics Shaped Statistics"

treatment – are in fact the same distribution.[27] Fisher formalized this test as the t-test which uses the t-statistic to quantify this discrepancy, and researchers across nearly every field to this day use p-values derived from t-statistics.

While in widespread use across several fields as the objective criteria for evaluating a causal hypothesis, Fisherian significance testing relies upon multiple subjective judgments in practice.

First, the threshold at which this statistic of discordance allows one to reject the null hypothesis is part of the pragmatics of experiments. While in early writings Fisher pointed to traditions in some fields of rejecting the null hypothesis at levels such as 5% or 1%, after criticism he stated that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas".

Second, the environmental conditions of the sampled population and the population in which a treatment will be ultimately used differ in ways known and unknown. An agricultural experiment may have been conducted in the presence of rain or no rain. An industrial experiment may have been conducted with one particular shift of workers.

Third, the probability that the null hypothesis is false actually says less than one might think about the probability that the alternate hypothesis – that there is an experimental effect – is true. Decisions are made based on substantive significance, which is the probability of an effect and the size of the effect. Jerzy Neyman and Egon Pearson attempted an experimental method that measured the probability of an effect, but it too requires the judgment as to what counts as an effect.

The identity crisis that these debates exposed was resolved through the publication of statistics textbooks beginning in the 1950s that integrated elements of the Fisher and the Neyman-Pearson methods into a recipe for significance testing that appealed to a mechanistic objectivity whose virtue was its freedom from human intervention. The consensus of several historians of statistics is that this was by design.[28]

Significance tests such as p-values, f-values and R^2 values today use arbitrary thresholds of statistical significance to infer the presence of an effect in any object of study or to select from competing statistical models of an effect and have been applied across dozens of fields from agriculture and psychology to medicine and industrial acceptance sampling. This application of statistics has been criticized as having impeded as much as it advanced technical progress, by

[27] The evaluation is based on the spread of the characteristic under experiment, and the discrepancy of the experimentally observed characteristic from the overall sample mean given the observed spread.

[28] "The need for personal judgment – for Fisher in the choice of model and test statistic; for Neyman and Pearson in the choice of a class of hypotheses and a rejection region; for the Bayesians in the choice of a prior probability – as well as the existence of alternative statistical conceptions, were ignored by most textbooks." Gerd Gigerenzer, et al, *The Empire of Chance*, (Cambridge University Press, 1989), P. 105

moving substantive domain expertise "out-of-the-loop" of work in favor of mechanical tests of statistical significance.[29]

This positioning of contemporary formalist statistics, as both objectively free of judgment and socially beneficial when applied, even though applied statistics requires judgment, while untenable has become dominant. Nonetheless the broader understanding of probability as a dialectical process for resolving ambiguity, in turn refining the judgment of domain experts, has continued to animate a strand of statistical work, from Shewhart and Deming to Tukey. When we advocate that AI workers be ethical by simply doing better AI, it is statisticians such as these to whom we appeal.

At Bell Labs, Walter Shewhart and his younger colleague Edwards Deming, were early critics of the displacement of workers' domain expertise with the use of arbitrary significance tests to judge quality of work. There were two distinct motivations for sampling-based inspection that reflected the internal tension within probability. In the context of the antagonistic relationship between labor and management, the most common approach was the use of acceptance sampling *tests*, which applied significance testing to reject work from vendors (in the case of inputs) or workers (in the case of outputs).

For Shewhart and Deming, the application of significance tests to control work was a relic of the 19th century elevation of exactness in industrial production, when the reality is that no industrial process can be reduced to a number. For them, variations in industrial output quality help workers better understand the processes themselves and inform new hypotheses into how to improve these processes.

For Deming, the development and teaching of statistics has been shaped excessively by the ideal of the elimination of judgment. Significance tests and related statistics for Deming point to no actual concepts or relationships in reality, and only conceal the full evidence given in the original data.[30] Deming advocates the presentation of evidential data as "distributions, scatter diagrams, and run charts to compare small groups and to detect trends". [31]

[29] "If you yourself deal in medicine or psychiatry or experimental psychology, …we would recommend that you focus on clinical significance. If you deal in complete life forms, environmental or ecological significance. If you deal in autopsies or crime or drugs, forensic or psychopharmacological significance. And so forth…An arbitrary and Fisherian notion of "statistical" significance should never occupy the center of scientific judgment." *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice and Lives*, by Deirdre McCloskey and Stephen Ziliak (Univ of Michigan, 2008), p. 20

[30] "Little advancement in the teaching of statistics is possible, and little hope for statistical methods to be useful in the frightful problems that face man today, until the literature and classroom be rid of terms so deadening to scientific enquiry as null hypothesis, population (in place of frame), true value, level of significance for comparison of treatments, representative sample. There is no true value of any concept that is measured. There may be, of course, an accepted operational definition (questionnaire, method of measurement) and an accepted value – accepted until it is replaced with one that is more acceptable to the experts in the subject matter". W. Edwards Deming, "On Probability As a Basis For Action", p. 151

[31] "Thus the prudent analyst will decide whether to calculate statistical intervals and stress the limitations of the resulting inferences, or to refrain from calculating such intervals under the belief that they may do more harm than good. In any case, these intervals may be secondary to the use of well-chosen statistical graphics to describe the data." "Assumptions for Statistical Inference", Gerald Hahn and William Meeker, The American Statistician, February 1993, Vol. 47, No 1, p. 10

This is particularly important given that the data upon which decisions are made – future crops, future workers – are always different than the data from which a sample is drawn to inform the decision.[32]

> There is no statistical method by which to extrapolate to longer usage of a drug beyond the period of test, nor to other patients, soils, climates, higher voltages, nor to other limits of severity outside the range studied….The gap beyond statistical inference can be filled in only by knowledge of the subject matter (economics, medicine, chemistry, engineering, psychology, agricultural science, etc.), which may take the formality of a model.[33]

Statistics most efficiently helps the subject matter expert to close this gap not with summary statistics of a large random sample, but with smaller samples stratified by time, location and environment that, with proper visualization, surfaces the underlying functional relationships. Thus in the majority of studies that lead to decisions on populations other than the sampled population, Deming called on statisticians to focus more on sampling and on visualization, both of which require intimate knowledge of the decision to be made, and less on statistical inference.

Deming's school of thought, while it has some adherents in corporate statistics, has generally been overshadowed by the more formalistic school of Neyman-Pearson, whose followers have argued for the use of tests of statistical significance as a mechanistic, universally valid approach to model selection that eliminates the subjectivity inherent in human decision making.

John Tukey, like Shewhard and Deming, drew a sharp distinction between mathematical statistics and data analysis in his classic 1962 paper, "The Future of Data Analysis"[34].

"Large parts of data analysis," wrote Tukey, "are inferential in the sample-to-population sense, but these are only parts, not the whole."[35] Tukey rejected the subjection of his work to mathematics implied by the term *applied statistics*, replacing it with the parallel field of *data analysis*.[36]

[32] While enumerative studies that informed decisions taken on the sampled population (e.g. whether to dispose of an order based on the weight of crates on a ship) called for randomized sampling and statistical inference, most decisions are taken on a different population than the one sampled and thus call for what Deming termed analytic studies.

[33] Ibid, p. 148

[34] Tukey, John, "The Future of Data Analysis," *The Annals of Mathematical Statistics*, 33, 1-67

[35] ibid, p. 2

[36] ibid, p. 3. To the extent that pieces of mathematical statistics fail to contribute, or are not intended to contribute, even by a long and tortuous chain, to the practice of data analysis, they must be judged as pieces of pure mathematics, and criticized according to its purest standards. Individual parts of mathematical statistics must look for their justification toward either data analysis or pure mathematics.

Exploratory data analysis, for Tukey, suggests new hypotheses by illuminating functional relationships not apparent in raw data, in contrast to confirmatory data analysis, which formally tests hypotheses. Modeling is a part of exploratory data analysis but is not the entirety.

Data analysis involves a healthy dose of judgement. “The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: ‘Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.’”[37] Like the classical probabilists, Tukey argued “The wise exercise of judgment can hardly help but the stimulate new theory”.[38]

Like Shewhart, Deming and many other statisticians throughout the 20th century, Tukey’s advocacy of a broader, dialectical notion of probability, and warnings of the limits of formalist probability, should be central to present-day appeals for AI workers to do ethical AI by simply doing better AI. By grafting external ethical concepts onto a narrowly instrumental AI, AI ethicists and ML fairness researchers endorse a formalist notion of probability that, like Shewhart, Deming and Tukey, we should resist.

[37] ibid, 13-14

[38] ibid, 10