Dialectical AI: Beyond the Accuracy-Ethics Trade-Off

By Ken Archer

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 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. 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] For dialectical probability, probabilistic models 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, in a virtuous dialectic through which scientific knowledge is developed within a domain or practice.

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

[1] Others have diagnosed the underlying ethical problem of AI as algorithmic formalism (Green 2019) and the formalism trap (Selbst 2019).

[2] Knight and Keynes in the early 20th century were notable for deepening our understanding of uncertainty as both doubt, or risk, and ambiguity.

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. 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, 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.

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

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

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 not separate from the concern for accuracy; these demands actually point to the essential role of social context in creating accurate algorithms.

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.[4]

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 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.[5] 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.

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

[5] 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.

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.*[6]

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 impact marginalized groups when used to design fair algorithms.[7] 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.

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

[7] 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.*

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.

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.

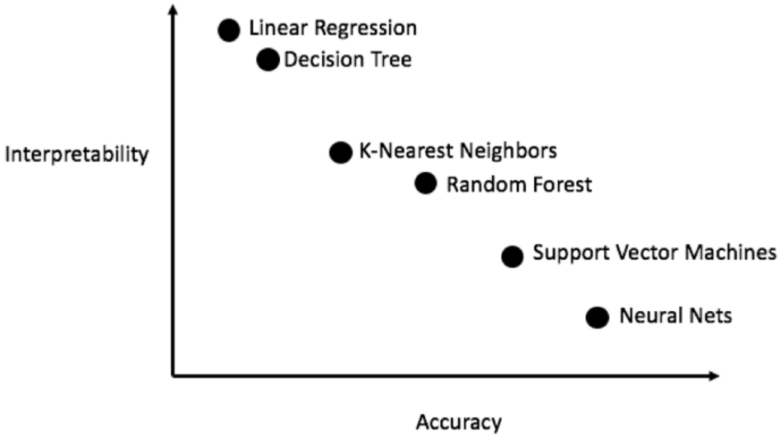
Explainability

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.[8]

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.



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

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

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.

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.[9]

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

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 (Green 2019) and the formalism trap (Selbst 2019). 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 approach to AI ethics 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, natural inclinations.

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.[10] This line of argument, too, buys

[9] 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.

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

into a formalistic understanding of how technology is built that is in fact contested within statistics and within other technical practices. Engineers don't need to yield their authority within their field of expertise to ethicists; they just need to be better engineers.

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 notion of technology as applied science espoused by educators conceals from engineers and statisticians.[11]

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.

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