Epistemic Therapy for Bias in Automated Decision Making

Thomas K. Gilbert
University of California, Berkeley
Berkeley, California 94703

Yonatan Mintz
Georgia Intititute of Technology
765 Ferst Dr NW
Atlanta, GA 30318

Abstract

Despite recent interest in both the critical and machine learning literature on "bias" in artificial intelligence (AI) systems, the nature of specific kinds of bias stemming from the interaction of machines, humans, and data remains ambiguous. Influenced by Gendler's work on human cognitive biases, we introduce the concept of *alief-discordant belief*, the tension between the intuitive moral dispositions of designers and the explicit representations generated by algorithms. Our discussion of *alief-discordant belief* diagnoses the various ethical concerns that arise when designing AI systems atop human biases. We furthermore codify the relationship between data, algorithms, and engineers as components of this cognitive discordance, comprising a novel epistemic framework for ethics in AI.

Introduction

As AI systems become pervasive, we have an interest in investigating their impact on fair decision-making. Both machine learning and social science researchers have confronted how the predictions made by these systems transpose human biases in certain contexts, adversely shaping decision-making in sensitive settings such as hiring (Gosh 2017), criminal justice (Angwin et al. 2016; Larson et al. 2016), and healthcare (Dwork et al. 2012; Aswani et al. 2016). To date this topic has been explored from both algorithmic (Hardt et al. 2016; Olfat and Aswani 2017; Bolukbasi et al. 2016) and positional perspectives (Breiman and others 2001; Dobbe et al. 2018; Mullainathan and Spiess 2017). However, the nature of the kinds of anthropomorphic biases that arise from the interaction between algorithms, humans, and data remains ambiguous.

We propose dissolving the concept of "anthropomorphic bias," and instead closely analyze its phenomenological content to better diagnose the epistemic and ethical challenges that may arise in AI system design. Since much of the discussion on anthropomorphic biases has been conducted through either anthropomorphizing AI models or relegating all responsibility for bias onto problematic data, our goal is to move away from these ambiguous analogies and confront the nuance of bias in automated decision-making.

Our discussion is motivated by several recent cases of AI system implementation and research that have already been flagged as problematic. We first examine systems where data was assumed to be representative of the real world but in fact encoded human biases that were later amplified by AI systems (Angwin et al. 2016; Reese 2016). We then consider cases in the machine learning literature where common notions of bias and equity are rationalized away by researchers using predictions provided by AI models (Corbett-Davies and Goel 2018; Wang and Kosinski 2017). Finally, we consider cases where researchers attempt to directly address anthropomorphic biases they perceive in their model prediction, leading to researchers instead encoding their own implicit biases into AI systems (Bolukbasi et al. 2016; Burns et al. 2018). These three motivating examples show that the nature of anthropomorphic bias in AI is paradoxical and counterintuitive, while the use of direct analogies to human bias actually erase the specific contexts that cause anthropomorphic bias to manifest in the first place.

In order to diagnose and work through these ambiguities, we provide two major contributions to the discussion of bias in AI systems. First, we characterize the kinds of biases that arise from complex interactions between engineers, algorithms, and data. To do this, we draw on previous work in the philosophy of mind literature that characterizes human bias as a set of aliefs (Gendler 2008), belief-like dispositions that also contain an affective component, and introduce the concept of *alief-discordant beliefs*. In our discussion, we demonstrate that these *alief-discordant beliefs* provide the framework necessary to understand the trade-off between the moral dispositions of system engineers and the explicit relations encoded by AI models. Second, we apply the notion of *alief-discordant beliefs* to craft design principles for AI in light of human biases. We conclude by illustrating how these principles can be used to diagnose the ethical concerns that motivate practitioners, performing a type of "epistemic therapy" for the benefit of the machine learning community.

Much as psychotherapy is used to confront subconscious issues through a process of dynamic interpersonal conversation, we propose "epistemic therapy" for automated decision-making as the process by which *alief-discordant beliefs* are identified and confronted by engineers. The design principles we outline in this paper support this "epistemic therapy" by aiding in the implementation of AI sys-

tems and navigating to avoid biased outcomes.

Notes on Technical Terminology

Given the interdisciplinary nature of our problem, we will first provide working definitions for technical terminology borrowed from philosophy of mind, machine learning, and the social sciences.

AI models: the actual computational function used to perform some kind of prediction task, in isolation from data as well as human data scientists and engineers.

AI system: the operation of AI in its full production context (including input pipeline and downstream decisions). In general, when there is no qualifier before AI, we make reference to this term.

Anthropomorphic bias: The attribution of cognitive biases, emotions, or intentions to AI systems. Note that this does not include forms of statistical bias or inferential limitations (e.g. sampling errors, misspecification, accuracy thresholds), and instead conflates the epistemic frames of humans and automated systems.

Epistemic ethics: A set of design principles that specify how AI systems should be operated and implemented. We distinguish it from classical ethical frameworks (e.g. utilitarianism) that present guidelines for human behavior alone; instead, epistemic ethics comprises rules for the deployment of systems in a context-appropriate manner.

Moral dispositions: The set of unconscious or semi-conscious behaviors, acquired through habituation and remaining active over the long-term, that guide moral action and personal character. While originating in humans, we will explore how they are unintentionally transposed into AI system and play a major role in anthropomorphic bias.

Techno-Societal Infrastructure: The dynamic relationship between AI models, data, and the individuals who design systems. It comprises all organizational resources responsible for an AI system's decision-making, rather than that system's mere operation. We refer interchangeably to data scientists, engineers, and machine learning researchers as the human component of this infrastructure.

Epistemic Ethics in Fair Machine Learning

The machine learning literature has explored predictive bias as a barrier to fair outcomes. In particular, many papers have suggested algorithmic means of curbing bias in predictions, proposing the notion of "optimal" fairness as solutions to optimization problems (Corbett-Davies and Goel 2018). Several variations of optimal fairness have been proposed, including equalizing prediction metrics (e.g. TPR, FPR, accuracy) across protected classes (Agarwal et al. 2018; Calders and Verwer 2010; Dwork et al. 2012; Hardt et al. 2016), and producing models whose predictions are independent of protected features (Burns et al. 2018; Bolukbasi et al. 2016; Zafar et al. 2015; Olfat and Aswani 2017). There have been some technical critiques leveled against these notions of optimal fairness (Corbett-Davies and Goel 2018; Mullainathan and Spiess 2017), as well as several impossibility results that show it is difficult to satisfy all of these notions of fairness simultaneously (Friedler, Scheidegger, and Venkatasubramanian 2016; Kleinberg, Mullainathan, and Raghavan 2016). We complement this literature by providing a framework of epistemic ethics that could be used to justify the necessary engineering design choices made in implementing certain methods of optimal fairness. Specifically, our framework resolves some of the skepticism in using these methods and addresses what technical knowledge and data engineers ought to have in order to manage and avoid anthropomorphic bias in AI systems they design.

A related stream of social science literature has discussed the nature of bias in machine learning (Barabas et al. 2017; Binns 2017; Mulligan et al. 2018; Dobbe et al. 2018), enumerating potential ethical concerns and discussing whether decision-making can be automated without compromising human dignity responsibility. In contrast, our discussion emphasizes how the *perception* of different kinds of bias in AI systems is responsible for their supposed immorality, which implies they are better understood within the realm of philosophy of mind than ethics proper. Of particular relevance to this paper is the discussion proposed by Binns (Binns 2017), which analyzes anthropomorphic bias by contrasting human mental states with the mechanisms of automated decision-making. We will present a complementary dissolution of anthropomorphic bias through the notion of *alief-discordant beliefs*, directly influenced by the work of Tamar Gendler on characterizing human biases (Gendler 2008).

Case Studies

We will first build intuition on how "anthropomorphic bias" can affect researchers and engineers through reference to prominent existing case studies. In particular, we consider cases where anthropomorphic bias may manifest in how 1) the values of engineers inform the techniques of AI system creation; 2) the epistemology of the data and encoding characterize the resulting AI system; 3) allocation of moral culpability labels interactions between data and system design throughout the deployment phase.

Each of these examples illustrates how anthropomorphic bias can be attributed to an underlying *discordance* between the distinctive epistemic frames of engineers, data, and system design. We discuss three major causes of this discordance, each with its own mixing of ethical culpability between data, algorithms, and engineers. These are: (i) AI systems that are deployed as vehicles for moral dispositions, either encoded by engineers or sedimented within data; (ii) engineers that outsource their own moral compasses to the labels generated by AI systems; (iii) dogmatic reconciliation of inherited dispositions and generated propositions that result in adverse effects.

AI systems as Vehicles for Moral Dispositions

First we consider cases where AI systems are deployed as vehicles for moral dispositions implicitly harbored by engineers or encoded in training data. Two famous cases of such discordance widely discussed in FAT literature include the COMPAS system which was used to predict criminal recidivism (Angwin et al. 2016; Larson et al. 2016), and Tay the Microsoft-deployed chat bot that was hijacked by white supremacists (Bass 2016; Price 2016; Reese 2016).

COMPAS The COMPAS system was initially developed to predict a recidivism risk score for arrested individuals and was deployed to several states (Angwin et al. 2016). However, even though explicit racial detail was not inserted into the system input, it was shown that COMPAS would predict lower risk scores for white individuals and higher scores for people of color. Moreover, analysis of the system output showed that the false positive rate for people of color was significantly higher than that of white individuals (Larson et al. 2016). From a technical perspective, it is generally agreed that the system produced biased output despite not taking in explicit racial data, since race is correlated with other features that were used as input (e.g. education, residence address, income) and labels used for training were obtained from historical arrest records that contain a documented inherent bias against certain communities (Larson et al. 2016).

In essence, the way the COMPAS system was designed and implemented transformed social biases extant in the justice system and reified them through automated classification scores. However, it is difficult to confer responsibility for this to a specific agent–while the data's encoded bias should have been made explicit by those who collected it, system engineers also failed to properly account for this bias in the data. What the engineers ought to have done in this particular case is *critically examine the source of the data* and any bias that it may convey, as well as *use a proper training technique* for their model that could account for the problematic predictions.

Microsoft Tay Tay was a chatbot developed by Microsoft research to interact with the greater public on social media and mimic the language patterns of a 19 year old girl(Bass 2016). While in closed company testing, it was reported that Tay was performing extremely well without significant incident. Within 24 hours of being deployed online, and to the surprise of the research team, Tay was re-tweeting white supremacist propaganda due to a loosely coordinated attack by certain forum users (Price 2016; Reese 2016).

Much like COMPAS, the deployment of Tay also suffered from becoming a passive vehicle for unquestioned moral associations. However, the associations in question are not necessarily the ones harbored by the data but by the engineers that designed the system. In particular, the engineers were not appropriately skeptical of the reaction of the internet community at large and assumed that they would behave in a similar manner to the corporate testers that interacted with Tay in house. The fallout from this situation could have been reduced had the engineers confronted their own underlying assumptions and properly designed the training methods of the model to not accept all input data equally.

Automatic Alief Falsification

Another cause of the underlying discordance can result from a 'skeptical' *overcorrection* based on the system pipeline. Two of the principle tenets of data science include the belief that all rational explanations (i.e. those that rely on mathematical and logical reasoning) are superior to all other forms of disposition generation, and that sufficient information about the state of the world can be extracted from data. This moral disposition is conveyed by both the recent work of Corbett-Davies and Goel on fairness measures (Corbett-Davies and Goel 2018) , and Wang and Kosinski's work on detecting sexual orientation using pictures (Wang and Kosinski 2017).

While we strive to identify ethical culpability in this section, we do not ascribe malicious intentions to the engineers that design AI systems. In particular, we believe both sets of authors that we discuss here did have good intentions when creating their work, but that their results are problematic when examined in a critical context. The key problematic aspect of these discordances is that instead of attempting to harmonize their own moral dispositions with the propositions generated by models, engineers are too eager to dispose of their initial presuppositions.

The Measure and Mismeasure of Fairness Corbett-Davies and Goel (Corbett-Davies and Goel 2018) have presented several leading methods of optimal fairness in the FATML literature and critique each family of methods to show how they may violate certain notions of fairness. The authors rely on working definitions of fairness from parts of the economics and legal literature and show how these can be incongruent with mathematical notions of fairness. When analyzing the notion of fairness through classification parity, the authors note that different populations of individuals will by nature have different means and variances that could account for lack of parity. As a real world example of this, the authors used the example of the COMPAS model we previously discussed. They note that since black individuals had higher recidivism rates than white individuals, this group was in fact accurately predicted as having a higher risk to society than whites.

Although the authors do note that this difference in rate is caused by both historic and system-specific factors, they argue that these may not be crucial to examine when making policy decisions (Corbett-Davies and Goel 2018). In particular, since the authors assume the prediction of individual risk is accurate, they claim that policy actions to ensure prediction parity between the populations would result in an unfairly harsher prediction rate against the white population. They further argue that such actions would harm the black population with an inappropriately low predicted rate of recidivism (Corbett-Davies and Goel 2018). Thus, the authors arrive at a conclusion that is in contrast to their initially stated assumptions ("demographic parity is important"), due to their failure to critically engage with the context of the data (namely its internal generation by a system with known bias (Larson et al. 2016)).

Detecting Sexual Orientation with AI A similar tension can be found in Wang and Kosinski's paper (Wang and Kosinski 2017), which describes an artificial neural network model that processes facial images to predict an individual's sexual orientation. In particular, the authors used face photos scraped from dating websites that they classified as heterosexual or homosexual using the user's dating profile, and showed that a model trained on these photos has good accuracy when predicting sexual orientation. The authors' main claim in this paper is that the predictive power of AI models

can be harnessed to encode complex patterns in facial features that could indicate an individual's sexual orientation.

This research has received a lot of backlash for suggesting a new form of digital physiognomy (Mattson 2017; Murphy 2017; Vincent 2017). While the researchers may have started with the assumption that physiognomy is pseudoscience, they readily discarded this in favor of the view generated by their model that facial features can predict sexual orientation. Kosinski himself has defended the study as revealing both the huge promise of big data as well as the risks due to loss of privacy (Resnick 2018). Meanwhile, critics have argued that this attempt to subject sexual orientation to objective measurement, while an interesting exercise in classification that reveals unexpected correlations with high accuracy, is erroneous as it fails to account for the subjectivity of social context (Gelman, Marrson, and Simpson 2018) and reifies social stereotypes(Miller 2018).

However, there is a danger of repeating the study's mistake by assuming that automated systems can only reify existing gender ontologies. It is, for example, possible that previously invisible correlations between bone structure and sexual preference really do exist, encouraging future work to explore and falsify new hypotheses. Meanwhile, a more critical analysis of the model might suggest that the contextual purpose of dating profile pictures is to broadcast sexual orientation to potential partners, rather than neutrally reflect how facial features predict sexual preferences. Since the engineers did not seriously consider this, they propagated a questionable conclusion based on the model output. We will later suggest interpreting such research studies as generating authentic discordances between our intuitions about the social world and novel beliefs about it that must be examined and deliberated, rather than summarily dismissed.

Dogmatic Reconciliation

Finally, we consider a class of discordances arising when the system engineers *do attempt* to harmonize their moral dispositions with system-generated propositions, but unfortunately use blunt methods to make systems comply with the former. The failure occurs when engineers, despite using values-based design, still encode their implicit biases into AI models through training and formulation instead of explicating and confronting their own assumptions. We consider two recent papers from the FATML literature that focus on the problem of biased predictions resulting from natural language processing: "Women also Snowboard" (Burns et al. 2018) and "Man is to Computer Programmer as Woman is to Homemaker" (Bolukbasi et al. 2016). In both of these papers, the authors seek to correct gender related bias in various downstream tasks that occurs using state-of-the-art word vector embeddings for natural language processing.

In "Women also Snowboard" (Burns et al. 2018) the authors address gender bias that occurs when developing AI systems for automatic image captioning. Specifically, the authors note that given certain contexts (e.g. sports equipment, computers, purses), image captioning systems tend to give incorrect predictions that fit common gender expression stereotypes. For instance, a captioning model might predict the caption of a picture of a woman snowboarding as a man snowboarding, since men are more associated with sports contexts. The solution that is introduced to curb this problem involves creating two classes of words ("male" vs. "female") and formulating a loss function to be used in model training that actually incentivizes confusion between these classes if insufficient evidence is found to make a gendered inference.

Likewise, in "Man is to Computer Programmer as Woman is to Homemaker" (Bolukbasi et al. 2016), the authors attempt to address gender bias that can be observed when performing analogy tasks using word vectors. While vector embedding in words generally yields useful semantic analogies (e.g. man is to king as woman is to queen), the authors note that certain problematic analogies are also picked up by these embeddings given certain corpora. To curtail these problematic analogies, the authors propose a method in which they compute a "Man to Woman" subspace of the embedded vector space, and formulate a way to reduce the the projection of non-gendered words on this subspace, thus removing much of the unintentional gender encoding that could have been contained by those words.

Both of these works confront a discordance between model and social understandings of gender, namely that it should not be informative for certain aspects of an individual (women can also be snowboarders and computer programmers), and attempt to harmonize this view with the generated propositions of these given models. That said, the solutions proposed by the authors do not directly correct for the notion that gender expression should be uninformative on certain predictions, and instead address the problem of women being underrepresented in data and should be predicted with equal probability to men in certain contexts.

This is a subtle distinction, but to illustrate it fully, we note that both solutions presented in these papers assume some kind of distinct "male" to "female" distinction and not more contextually-nuanced forms of gender identity. Essentially, by attempting to produce a solution to the systems discordant moral claims, the authors have hard-coded a cisgender understanding of human sexuality into the models. The authors thus do not directly engage with the *root discordance* they seek to address, and instead provide a potentially problematic stop-gap solution that reacts only to their morally-charged dispositions. One way of mitigating these effects would have been a more thorough examination of the social-scientific literature on gender expression, and broadening the gender diversity of the research teams involved.

The Context of Anthropomorphic Bias

The questions surrounding anthropomorphic bias raised by our prior discussion–is it always traceable to some original context of human bias? is it original to the statistical compromises that accompany automated decision-making? can it possibly be avoided entirely?–are considerable and demand a deeper philosophical analysis. In this section we proceed in two parts.

Gendler on Belief-Discordant Alief

Some form of "bias", however it is defined, is inevitable when any small team of humans derives actuarial interventions for broad populations. Trade-offs between accuracy

and variance or false positive vs. false negative rates have been a hallmark of statistical inference since the discipline's birth, and while AI has considerably increased the scale and speed of such inferences in deployment, they have not fundamentally changed the rules of this game. Instead we should aim for a principled trade-off between the limits of inference, given messy data sets, imperfect model choices, or limited training time. But what might such a principled trade-off look like for anthropomorphic bias, which combines affect-laden human intuitions with machines' capacity for semantically-arbitrary classification?

For conceptual guidance, we appeal to Gendler's (Gendler 2008) work on the complex and codependent relationship between belief and alief, which defines the latter as "a mental state with associatively-linked content that is representational, affective and behavioral, and that is activated – consciously or nonconsciously – by features of the subject's internal or ambient environment. Aliefs may be either occurrent or dispositional."

Gendler illustrates this phenomenon through the concept of *belief-discordant alief*, which accounts for a diverse set of scenarios; for example, people who are afraid of walking on an open skywalk despite its structural safety, people who won't touch an object for fear of "cooties," automatically reaching for one's wallet when one knows one left it at home, and being afraid of something on the screen in a movie theater. Belief-discordant alief is the triggering of affective response patterns and automatic motor routines opposed to "explicit, conscious, vivid, occurrent belief" (Gendler 2008). That is, it arises when we enter a situation that triggers us into a cognitive state that counteracts what our 'better' judgment knows not to be the case.

There is some semantic risk in applying such technical philosophical concepts to a problem as provocative and wide-ranging as anthropomorphic bias, which is already the subject of a rapidly growing empirical literature. However, we feel Gendler's language is not just relevant but necessary for diagnosing the problem in machine learning, for two reasons. First, Gendler's examples and qualifiers succeed in contrasting belief with alief by defining the former in a strongly *computational* sense: belief is an explicit proposition whose content is discrete (not associative), is universally held (not situationally triggered), and refers conclusively to external reality (not emotions or habits of mind). We shall see that these technical descriptors of belief and alief are extremely useful for diagnosing the specific epistemic tensions within AI systems and the moral dispositions of those who design or interpret them.

Second, Gendler's terminology emphasizes the *discordance* between different kinds of bias as the source of the real problem, not implicit bias in isolation. She holds, following Hume, that the hallmark of alief is a kind of association by which semantic, emotional, and behavioral dimensions are crystallized over time. In belief-discordant alief, there is something about an environment's psychological effects that trigger one to automatically respond in a way opposed to one's beliefs about it. The associative content of alief is highly arbitrary, just as one's explicit beliefs may be fundamentally prejudiced. This is key for grasping the confusions surrounding anthropomorphic bias in machine learning: AI systems aren't conscious, yet classify social artifacts much as we do; system designers strive for formal accuracy, yet display strong moral affect in response to automated claims. The biases of both systems and designers play a role in generating the discordant environments in which a machine's classifications feel inappropriate or morally wrong, and we can make sense of this by recognizing the ontological primacy of belief-alief discordance over the isolated prejudices of humans and machines.

Alief-Discordant Beliefs in Machine Learning

To properly apply Gendler's insights, we propose the concept of *alief-discordant belief* to describe the origin, form, and consequence of anthropomorphic bias in automated systems. When deployed, these systems (e.g. image captioning) transpose human-generated forms of alief by computationally remaking the context within which our aliefs typically operate. In Gendler's terms, they generate beliefs that violate the habitual associations between semantic meaning (e.g. these variables are related given a specific parameter space) and moral dispositions (snowboarding is something anyone can do), producing a visceral reaction from the designers ("women also snowboard") that demands reconciliation (Burns et al. 2018). Consequently, the "bias" of automated systems refers to the uncanny semantic associations that arise from applying our aliefs to a purely data-driven setting, violating the contextual ties between disposition, affect, and representation that underpin our aliefs.

When confronted with examples of an image classifier that offend us, we may have an automatic affective response that counteracts the beliefs that are either encoded into the algorithm's learning procedure (the belief that the classifier can learn semantic associations in an objective, impartial manner and arrive at ground truth) or generated by it (only men snowboard, certain faces are gay). In other words, the classifier crystallizes propositions about its own learning procedure as well as its predictive outputs, either of which can give rise to alief-discordant belief.

Although such a classifier is just a computational function, there is a tendency for designers to anthropomorphize it as if it had autonomous beliefs, leading to a search for where these beliefs come from or who is to blame. This can leave engineers in the position of apologizing for the very data that is needed for the model learn anything useful (much like the case of MS Tay learning English for Nazi posts).

We suggest interpreting such systems as *automatic belief generators* that compel us to reinterpret our own aliefs in the deployment context. Rather than claiming an algorithm is "biased," we should confront the tension it creates between beliefs and aliefs we have about the social world. We must avoid compounding this tension by manually patching in solutions to our most violated or dogmatically-held aliefs (e.g. we do calibration to make a classifier generate all outcomes independently of protected attributes). Instead, we must address the *discordance* that is making us feel uncomfortable (such as in the cases as (Burns et al. 2018) and (**?**))).

Rather than blaming the data or its labelers as biased, system designers are responsible for sorting out these discor-

dances as they arise by harmonizing the generated beliefs with their own newly-challenged, inherited aliefs. The goal is thus not to try to make automated systems unbiased, but to interrogate the beliefs generated, the procedure for that generation, and the relation of both to our own aliefs. These factors can be made to map onto specific components of the AI's socio-technical infrastructure: the system engineer, the chosen model and training methods, and the data used. The onus is therefore on figuring out where in this pipeline our aliefs are being violated, how each contributes to this violation, and which of these components is most responsible.

Towards a Techno-Societal Infrastructure

What is the relationship between the structure of the machine learning pipeline and the generation of alief-discordant belief? Here we argue that alief-discordant beliefs, while inevitable, are managed most easily and readily if each step of the pipeline is designed to maintain one analytic component of this discordance. In other words, all layers–which we divide below into data, model training, and engineers–must be acknowledged as co-responsible for the discordance and play distinctive roles in its generation. Once these roles are analytically distinguished, it is much easier to identify the best practices for each so that the source(s) of anthropomorphic bias can be readily diagnosed.

Our goal is therefore to trace the alief-belief concordance in datasets all the way to the inevitable discordances produced through AI deployment. Pre-reflective moral dispositions necessarily inform the application of this pipeline through continuous attention to context, but the mechanics themselves are tied to conditions of explicit knowledge representation in the form of system-generated propositions. A significant component of "fair" machine learning is the integrity and documentation of this techno-societal pipeline, such that bias can be managed well beneath the psychological threshold of moral outrage that has regrettably defined the public reception of prominent case studies (see for example (Snow 2018)). This implies that much of the burden for moral responsibility gets shifted from the data to model training and finally to engineers–if practitioners find themselves ignorant of the context of what they are working on, it is inappropriate to shift blame onto your tools.

Data

The collection of data about human subjects requires a compromise between alief (whatever moral compulsion(s) was felt in the leadup to collecting it) and belief (whatever data structure and type was determined to best represent assumptions about reality). There is always a moral context that informs the data content, just as there is an explicit frame of representation that accounts for its form (recall the cases of COMPAS and Tay). This original compromise is often invisible and unacknowledged in a given dataset, as social categories are mutable and tied to the historical perspectives of those who produced and tabulated the data. While troubling, the main problem here is how to deal with this crystallization to ensure it can be *accounted for* and *is traceable*. Here the following guidelines are necessary:

Data as context-specific: data must be documented with the original priorities of those who collected it, and the relevant case law that informed its collection. These "datasheets for datasets" help ensure that transparency and accountability are baked into the pipeline from the start (Gebru et al. 2018). For example in the case of university research, the data should be tied to its original IRB protocol, including the motivations for the study. This will provide a traceable baseline for the alief(s) lying behind the original data even after it has been used to generate classification regimes.

Data as explicit: data should be annotated so that the assumptions behind its collection are clearly documented, rather than left implicit. This implies that where sample sizes are unevenly dispersed across subpopulations, the data gatherers provide an account of why they felt it was still representative of the underlying social reality or at least indicative. This is meant to account for the beliefs or prior evidence that informed why the data was collected and organized in a certain manner, rather than another.

Data as contestable: data should be publicly available so that its alief-belief crystallization, however messy or regrettable, can be challenged by those most subject to its labels or classification. As it embodies a representation of social reality, data is directly contestable as a ground source of belief propositions, unlike engineers (see below). This affirms that the alief-belief matrix behind its collection is a product of compromise worthy of continual reflection and deliberation.

Where any of these principles are not obeyed (as is likely), responsibility falls on model selection and training.

Model and Training

As a machine learning model is being trained, data–and with it a complex web of social relations, moral drives, and unstated representational axioms–are reified as generated beliefs about the macro-environment in question. This is the great benefit and cost of machine learning at scale, to wring more intuition out of a dataset than existed anywhere in the minds of those who produced it. Consequently, the main problem is to ensure the generated beliefs *accord with* the inherited beliefs that defined that data, or at least the existing beliefs of system engineers:

Models as interpretable: models should be easy to understand by qualified humans so that classifications have a clear semantic context, requiring *explainability*. If this is not the case, as can happen in deep learning (Girshick et al. 2014), it will make the work of identifying bias (and particularly its causes) difficult, because the model itself cannot be consulted to resolve the discrepancy or point to likely solutions. This may leave a significant gulf between the beliefs of data scientists and the generated beliefs of the model that will pop up as a discordance once system engineers confront it later in the pipeline. This "discordance slack" should instead be reined in as early as possible.

Models as intuitive: model assumptions should be documented so they are easily altered, without an unnecessary amount of work going into why these assumptions were chosen. This is necessary to ensure that the *belief generation* of the model is kept distinct from the *belief testing* of its trainers, rather than the two becoming conflated. In other words,

trainers should imagine, tabulate, and try to justify all possible model choices before converging on one, and should provide documentation of this deliberative process.

Models as corrigible: assumptions should be transparent and available to the wider machine learning community so that they can be challenged. More specifically, machine learning specialists who were not part of the training process and were not familiar with the original data should be able to interrogate, question, or reject a model's assumptions if they supply sufficient grounds for doing so. This is a redundancy check against the psychological biases of the model trainers and helps bolster the accordance between humans' and the model's beliefs.

Where these principles are disobeyed, which for some models (e.g. unsupervised learning) is inevitable, engineers are responsible.

Engineers

Engineers are the ultimate source of aliefs that are discordant with the model's generated beliefs. As those who are using the model to classify new data in contexts other than what the model was trained on, they will often bear the responsibility for its failures and specifically for the discordance between their own moral dispositions and the model's classifications. To review what has been stated already, where the data is dirty and the model is a black box, alief-discordant belief is almost inevitable. This is the scenario we are trying to anticipate by making the discordance *manageable*.

Discordance as discoverable: engineers should be trained about psychological bias to better identify discordances where they are subtle or hidden in edge cases for the model in question. This includes an awareness of their teammates' professional background, so that double-blind system checks can be conducted to maximize the model's robustness given the prejudices and assumptions of those auditing it. The goal is to make note of discordance before the public discovers it in deployment, which could cause social harm and also make the model less trustworthy.

Discordance as tractable: engineers should be expertly informed about the likely problems with data and model training, i.e. have a general grasp of the pipeline and its context to better confront discordances when or if they arise. This includes educational programming about historical cases of model bias, what forms of bias are most common, and how these biases typically arise for either the choice of model or specific dataset. The goal here is to know how to manage the discordance, rather than simply to flag it.

Discordance as contextual: engineers should be trained to maintain healthy work environments, have access to legal consultation, and cultivate emotional intelligence to better process discordances when they themselves feel them. They should also maintain awareness that, however morally charged, unfair classifications are survivable and must know how to convey them to different qualified professionals (e.g. peers, lawyers, researchers) to mitigate the damage they may cause once deployed. For this reason, engineers are not contestable in the same way that data is, as it is merely their *aliefs* (which are non-propositional) that are responsible for the discordance, not machine-generated or data-encoded *beliefs* (which are propositional). That being said, thanks to Three Mile Island and other man-made disasters, we have learned that fear-mongering, societal distrust, and lasting damage can be avoided if engineers respond appropriately to a crisis rather than misrepresent the nature of the problem either to themselves or the public.

Where these principles are disobeyed, engineers are morally (and perhaps legally) responsible for resulting harm.

Conclusion

We have suggested *alief-discordant belief* as a term that avoids the shoals of anthropomorphic bias in an automated context. Epistemically, *alief-discordant belief* accounts for the subtle ways in which human cognitive bias enters a machine learning pipeline, first through the dataset, then through model and training specification, finally in the temperament and disposition of system engineers. Ethically, we have suggested that machine learning practitioners should work to maintain the integrity of this pipeline so that *alief-discordant belief*, once generated by engineers interacting with the model, is manageable with respect to the actual stakes of the social context in question. This implies that part of the wider project of realizing fairness through machine learning is for engineers to interpret themselves as part of this context, which includes the wider machine learning community as well as potentially-vulnerable populations of protected social categories. In light of this therapy, negotiating the discord between human aliefs and machine-generated beliefs may depend on crafting a wholly new context for automated decision-making, in which we get better at designing machines that supply us with beliefs that we are more critically prepared to adopt.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*.

Aswani, A.; Kaminsky, P.; Mintz, Y.; Flowers, E.; and Fukuoka, Y. 2016. Behavioral modeling in weight loss interventions.

Barabas, C.; Dinakar, K.; Virza, J. I.; Zittrain, J.; et al. 2017. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv preprint arXiv:1712.08238*.

Bass, D. 2016. Clippy's back: The future of microsoft is chatbots. bloomberg businessweek.

Binns, R. 2017. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357.

Breiman, L., et al. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16(3):199–231.

Burns, K.; Hendricks, L. A.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*.

Calders, T., and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Dobbe, R.; Dean, S.; Gilbert, T.; and Kohli, N. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumeé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gelman, A.; Marrson, G.; and Simpson, D. 2018. Gaydar and the fallacy of objective measurement. *Unpublished manuscript. Retrieved from http://www. stat. columbia. edu/˜ gelman/research/unpublished/gaydar2. pdf*.

Gendler, T. S. 2008. Alief and belief. *The Journal of philosophy* 105(10):634–663.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Gosh, D. 2017. Ai is the future of hiring, but it's far from immune to bias. *Quartz*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)* 9.

Mattson, G. 2017. Artificial intelligence discovers gayface. sigh. `https://greggormattson.com/2017/09/09/artificial-intelligence-discovers-gayface/`.

Miller, A. E. 2018. Searching for gaydar: Blind spots in the study of sexual orientation perception. *Psychology & Sexuality* 1–16.

Mullainathan, S., and Spiess, J. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31(2):87–106.

Mulligan, D.; Elazari, A.; Burrell, J.; and Kluttz, D. 2018. Afog workshop panel 2: Automated decision-making is imperfect, but it's arguably an improvement over biased human decision-making. Technical report, University of California, Berkeley.

Murphy, H. 2017. Why stanford researchers tried to create a 'gaydar' machine. *New York Times*.

Olfat, M., and Aswani, A. 2017. Spectral algorithms for computing fair support vector machines. *arXiv preprint arXiv:1710.05895*.

Price, R. 2016. Microsoft is deleting its ai chatbot's incredibly racist tweets. *Business Insider*.

Reese, H. 2016. Why microsoft's tay ai bot went wrong. *Tech Republic*.

Resnick, B. 2018. This psychologists gaydar research makes us uncomfortable. thats the point. *Vox*.

Snow, J. 2018. Google photos still has a problem with gorillas.

Vincent, J. 2017. The invention of ai 'gaydar' could be the start of something much worse. *The Verge*.

Wang, Y., and Kosinski, M. 2017. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.