In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017) forthcoming*.

Algorithmic Bias in Autonomous Systems

David Danks[1,2] **& Alex John London**[1,3]
[1]-Department of Philosophy; [2]-Department of Psychology; [3]-Center for Ethics and Policy
Carnegie Mellon University, Pittsburgh, USA
{ddanks, ajlondon}@andrew.cmu.edu

Abstract

Algorithms play a key role in the functioning of autonomous systems, and so concerns have periodically been raised about the possibility of *algorithmic bias*. However, debates in this area have been hampered by different meanings and uses of the term, "bias." It is sometimes used as a purely descriptive term, sometimes as a pejorative term, and such variations can promote confusion and hamper discussions about when and how to respond to algorithmic bias. In this paper, we first provide a taxonomy of different types and sources of algorithmic bias, with a focus on their different impacts on the proper functioning of autonomous systems. We then use this taxonomy to distinguish between algorithmic biases that are neutral or unobjectionable, and those that are problematic in some way and require a response. In some cases, there are technological or algorithmic adjustments that developers can use to compensate for problematic bias. In other cases, however, responses require adjustments by the agent, whether human or autonomous system, who uses the results of the algorithm. There is no "one size fits all" solution to algorithmic bias.

1 Introduction

Algorithms play a critical role in all computational systems, and particularly autonomous ones. In many ways, algorithms—whether those implemented in the autonomous system itself, or those used for its learning and training—constitute the "mind" of the autonomous system. Autonomy requires capabilities to adapt and respond to novel, often ill-defined, environments and contexts. And while hardware and other software components are obviously important, algorithms are the key to these abilities. In particular, we focus here on learning, context detection, and adaptation algorithms for autonomous systems, regardless of whether the algorithms are employed in training and development, or in real-time system activity, or in both regimes.

In many cases, we have turned towards autonomous systems precisely because they do not have some of the flaws or shortcomings that we humans have. For example, a self-driving vehicle cannot fall asleep at the wheel, or become distracted by background music. If autonomous systems are to be better versions of us (at least, for some tasks), then we should plausibly aspire to use the most unbiased algorithms that we can.

Despite this aspiration, several high-profile cases have prompted a growing debate about the possibility, or perhaps even inevitability, of *algorithmic bias*: roughly, the worry that an algorithm is, in some sense, not merely a neutral transformer of data or extractor of information. The issue of algorithmic bias has garnered increasing attention in the popular press and public discussions of technology, including widespread concerns about "bias" (of one form or another) in Google searches, Facebook feeds, applications such as FaceApp, and other algorithmic systems. Moreover, there is a rapidly growing scholarly literature about algorithmic biases, including technological techniques to try to mitigate it (e.g., [Barocas and Selbst, 2016; Garcia, 2016; Kirkpatrick, 2016; Pedreschi, *et al*., 2008]).

The possibility of algorithmic bias is particularly worrisome for autonomous or semi-autonomous systems, as these need not involve a human "in the loop" (either active or passive) who can detect and compensate for biases in the algorithm or model. In fact, as systems become more complicated and their workings more inscrutable to users, it may become increasingly difficult to understand how autonomous systems arrive at their decisions. To the extent that bias is determined by the process of decision making and not solely by outcomes, this inscrutability may challenge the very notion of human monitoring for bias. And so while autonomous systems might be regarded as neutral or impartial, they could nonetheless employ biased (in some sense) algorithms that do significant harm that goes unnoticed and uncorrected, perhaps until it is too late.

As an example of such concerns (though not involving an autonomous system), there have been several high-profile demonstrations of systematic racial bias in algorithms used to predict recidivism risk (i.e., the likelihood that an individual convicted of a crime will commit another crime in the future). These prediction algorithms have been touted as "more objective" or "fairer," and yet they seemingly exhibit quite systematic biases against particular racial

groups, perhaps because they encode broader systematic issues [ProPublica, 2016].

While we agree that there are very real worries here, we also contend that many distinct issues have been unhelpfully lumped together under the title of "algorithmic bias." Public discussions of algorithmic bias currently conflate many different types, sources, and impacts of biases, with the net result that the term has little coherent content. For example, algorithmic bias in the recidivism prediction case involves different possible statistical, ethical, and legal biases, all entering in different places and in different ways. There is no coherent notion of "algorithmic bias" in this case, or in most others. And correspondingly, there is little reason to think that there is one consistent or reliable response to these myriad possible biases. Proper mitigation measures, and whether we should respond at all, depend deeply on the nature and source of the bias, as well as the norms and values to which the performance of the system in question must be accountable.

This paper is an effort to provide some structure and clarity about concepts of, concerns about, and responses to algorithmic bias. Section 2 provides a taxonomy of different notions that one might have in mind with the term 'algorithmic bias'. We then turn in Section 3 to consider the issue of appropriate responses to algorithmic bias: when is a response warranted, and what form(s) should it take? And throughout, we focus on algorithmic bias in autonomous systems; this is obviously not the only context in which we can face significant or harmful algorithmic bias, but it is a particularly important one, given the decision-making power accorded to an autonomous system.

2 A Taxonomy of Algorithmic Bias

The word 'bias' often has a negative connotation in the English language; bias is something to be avoided, or that is necessarily problematic. In contrast, we understand the term in an older, more neutral way: 'bias' simply refers to deviation from a standard. Thus, we can have statistical bias in which an estimate deviates from a statistical standard (e.g., the true population value); moral bias in which a judgment deviates from a moral norm; and similarly for regulatory or legal bias, social bias, psychological bias, and others. More generally, there are many types of bias depending on the type of standard being used.

Crucially, the very same thing can be biased according to one standard, but not according to another. For example, many professions exhibit gender disparities, as in aerospace engineering (7.8% women) or speech & language pathology (97% women) [https://www.bls.gov/cps/cpsaat11.htm]. These professions clearly exhibit statistical biases relative to the overall population, as there are deviations from the population-level statistics. Such statistical biases are often used as proxies to identify moral biases; in this case, the underrepresentation of women in aerospace engineering may raise questions about unobserved, morally problematic structures working to the disadvantage of women in this area. Similarly, the over-representation of women in speech and language pathology may represent a moral bias, depending on additional information about the proper moral baseline or standard for making such moral assessments.[1]

These observations are relatively uncontroversial on their own, but they already present a problem for the notion of "algorithmic bias": there are multiple, different categories of bias—each of which could be further subdivided—that are often treated as equally problematic or significant, even though not all forms of bias are on a par. Some may be deeply problematic deviations from a standard, while others may be valuable components of a reliable and ethically desirable system. Moreover, these issues often cannot be resolved in a purely technological manner, as they involve value-laden questions such as what the distribution of employment opportunities *ought* to be (independently of what it actually empirically is), and what factors ought and ought not to influence a person's employment prospects.

Equally importantly, these different biases for algorithms can arise from many different sources. We thus turn to the task of disentangling different ways in which an algorithm can come to be biased. Although these different sources sometimes blur together, the taxonomy we provide creates a richer space within which to assess whether a particular bias merits a response, and, if so, what sort of corrective or mitigation measures might be implemented.

2.1 Training Data Bias

One route to algorithmic bias is through deviations in the training or input data provided to the algorithm. Algorithms are trained or learn for particular uses or tasks (e.g., for the population from which samples are drawn). The input data that are used, however, can be biased in one or another way, and thereby lead to biased responses for those tasks. In particular, a "neutral" learning algorithm (in whatever sense of that term one wants) can yield a model that strongly deviates from the actual population statistics, or from a morally justifiable type of model, simply because the input or training data is biased in some way. Moreover, this type of algorithmic bias (again, whether statistical, moral, legal, or other) can be quite subtle or hidden, as developers often do not publicly disclose the precise data used for training the autonomous system. If we only see the final learned model or its behavior, then we might not even be aware, while using the algorithm for its intended purpose, that biased data were used.

As an uncontroversial example, consider the development of an autonomous vehicle, such as a self-driving car, and suppose that the vehicle is intended for use throughout the United States. If the vehicle's training data and information

[1] Statistical information may be asymetrically informative in these professions as there is a history of gender-based workplace discrimination against women, but not men. We might reasonably think that deviations from population statistics are more likely to reflect gender-based discrimination when they disadvantage women. At the same time, it is certainly possible that there are morally problematic factors that disadvantage men in the field of speech & language pathology. One needs a well-defined moral standard to make such assessments, as well as particular empirical facts that likely go beyond just unequal gender representation.

come entirely or mostly from one location or city (e.g., the Google cars in Mountain View, or the Uber cars in Pittsburgh) and we use a relatively "neutral" learning algorithm, then the resulting models will undoubtedly be biased relative to the intended purpose and scope. For example, the use of training data from only Pittsburgh could lead the self-driving vehicle to learn regional norms or customs, rather than patterns that apply across the intended context of driving throughout the U.S. More generally, we would have significant training data bias, since our data come only from a small part of the world. As a result, significant problems could result if this autonomous vehicle were placed (without supervision) in the broader, intended contexts of use.

This case makes vivid the importance of being clear about the relevant standard against which we judge an algorithm (or algorithm output) to be biased, particularly when that standard is determined by the intended uses of the algorithms or resulting models. Relative to the standard of statistical distribution for Pittsburgh, for example, the self-driving vehicle might well exhibit *no* algorithmic bias (or at least, no statistical bias due to training data). In this case, the algorithmic bias due to training data obtains only relative to a different standard—namely, the statistical distribution over a much larger geographic area.

This example centers on bias relative to a statistical standard, but training data bias can also lead to algorithmic bias relative to a moral or normative standard. For example, suppose that we are training a prediction algorithm that will subsequently be used to make healthcare allocation decisions in a population. This algorithm will causally influence the future population, and so we might think it important to ensure that it does not maintain a morally problematic status quo. Because the relevant moral standard (about the population) is different from the current empirical facts, we might actually choose to train the algorithm using data that reflects the *desired* statistical distribution. That is, there might be cases in which we deliberately use biased training data, thereby yielding algorithmic bias relative to a statistical standard, precisely so the system will be algorithmically *unbiased* relative to a moral standard.

2.2 Algorithmic Focus Bias

A second, related route to algorithmic bias is through differential usage of information in the input or training data. We often believe that an algorithm ought not use certain types of information, whether for statistical, moral, legal, or other reasons, even if those variables are available. The obvious case is the use of morally irrelevant categories to make morally relevant judgments, though things can get quite complicated when the morally irrelevant category is statistically informative about, but not constitutive of, some other, morally relevant category (see Section 3 below). In these cases, a source of algorithmic bias relative to a statistical standard can be the deliberate non-use of certain information, as that can lead to an unbiased algorithm relative to a moral standard.

A more neutral instance in which algorithmic focus can lead to bias arises in the use of legally protected information in certain types of decision-making. An otherwise "neutral" learning algorithm might nonetheless exhibit algorithmic bias (relative to a legal standard) due to a biased focus if it is provided input variables that are not legally permitted to be used for certain types of predictions or judgments. The algorithm will deviate from the legal standard, even if it is plausibly statistically unbiased (assuming unbiased training data). That is, we can have a case with a "forced choice" between two types of algorithmic bias: one relative to a legal standard through the use of information that violates a legal standard, versus one relative to a statistical standard by ignoring statistically relevant information in the input data.

2.3 Algorithmic Processing Bias

A third source of algorithmic bias arises when the algorithm itself is biased in various ways. The most obvious instance of algorithmic processing bias is the use of a statistically biased estimator in the algorithm. Of course, there might be good reasons to use a statistically biased estimator; most notably, it might exhibit significant reduced variance on small sample sizes (i.e., the bias-variance tradeoff [Geman *et al*., 1992]), and thereby greatly increase reliability and robustness in future uses. That is, we might embrace algorithmic processing as a bias source in order to mitigate training data as a source of bias (Section 2.1).

In fact, many, perhaps even most, cases of bias due to algorithmic processing arise through deliberate choice: we consciously choose to use a "biased" (in some sense) algorithm in order to mitigate or compensate for other types of biases. For example, if one is concerned about the biasing impacts of training data, then many algorithms provide smoothing or regularization parameters that help to reduce the possibility of overfitting noisy or anomalous input data. While this choice might be absolutely correct in terms of future performance, it is nonetheless a source of algorithmic bias, as our learning algorithm is not neutral (in a statistical sense). As we noted earlier, not all biases—algorithmic or otherwise—are bad.

In the context of autonomous systems, algorithmic processing is arguably a widespread source of bias, precisely because of the importance of ensuring robustness in our algorithms. It also arises in cases such as "ethical governors" that alter the output of the learning algorithm so that the autonomous system is more likely to make ethical choices (in some sense), even at the cost of reducing the likelihood of success on non-moral mission-oriented criteria. For example, an autonomous weapons system might be provided with an ethical regulator that will not allow it to fire at perceived enemy combatants if they are near a UNESCO protected historical site. The processing of the algorithm is statistically biased in the sense that its judgments or decisions deviate from what a "neutral" algorithm might have done, but it helps ensure the system conforms to important moral norms [Arkin *et al*., 2012]. These ethical modules bias the algorithms in important ways, though not negatively.

2.4 Transfer Context Bias

The previous three routes to algorithmic bias centered on technical or computational matters. In contrast, the last two arise from inappropriate uses or deployment of the algorithms and autonomous systems. As noted earlier, we deploy algorithms for particular uses or purposes, and in particular contexts of operation, even if those are often not explicitly stated. However, when the algorithm or resulting model is employed outside of those contexts, then it will not necessarily perform according to appropriate standards, whether statistical, moral, or legal. Of course, there is a sense in which this is arguably *user* bias, not algorithmic bias. Nonetheless, we contend that many cases that get labeled as "algorithmic bias" are actually due to unwarranted application or extension of an algorithm outside of its intended contexts.

For example, consider the earlier discussion of self-driving vehicles intended for use throughout the U.S. These autonomous systems would clearly perform in a biased (in the negative sense) manner if they were deployed in, say, the United Kingdom, since people drive on the left-hand side of the road there. Moreover, this biased performance arises from inappropriate use outside of intended contexts. A more subtle example of transfer context bias could arise in translating a healthcare algorithm or autonomous system from a research hospital to a rural clinic. Almost certainly, the system would have significant algorithmic bias relative to a statistical standard, as the transfer context likely has quite different characteristics. This statistical bias could also be a moral bias if, say, the autonomous system assumed that the same level of resources were available, and so made morally flawed healthcare resource allocation decisions.

There is a fine line between transfer context and training data sources of bias. For example, if the self-driving vehicle had been intended for worldwide driving, then we would arguably have a training data source of bias, not a transfer context bias. There is, however, an important general difference between (a) learning from biased data about the intended contexts of operation (training data source); and (b) inappropriately using an algorithm outside of its intended contexts of operation (transfer context source).

2.5 Interpretation Bias

A final source of algorithmic bias is misinterpretation of the algorithm's outputs or functioning by the user, or by the broader autonomous system within which the algorithm functions. While this bias might be characterized simply as user error, the situation is often more complex than this. In particular, interpretation bias represents a mismatch, even within the intended context of operation, between (i) the information an algorithm produces; and (ii) the information requirements of the user or system that uses that output. Moreover, there is widespread potential for this kind of informational mismatch, since developers are rarely able to fully specify the exact semantic content (in all contexts) of their algorithms or models. Systems that take this output as input can thus easily be misdirected by spurious or unreliable features of that information.

As a simple example, consider the use of regression analyses to generate causal or policy predictions for subsequent decision-making. Standard regression methods yield non-zero coefficients for any input variable that is statistically associated with the target, conditional on the other input variables. Thus, effects will typically have non-zero regression coefficients with respect to their causes, even though the causal flow goes in the opposite direction. As a result, non-zero regression coefficients should not be interpreted as indicating degree of causal strength, even though this practice is quite common in certain scientific domains. Similarly biased judgments about causal structure or strength (i.e., that deviate from the actual causal structure in the world) can easily be misused in biased ways by autonomous systems.

As a more practical example, consider an autonomous monitoring system that makes decisions about how to shift its surveillance resources to track the most relevant targets. Such a system presumably has one or more algorithms for inferring the "surveillance value" of different individuals. However, there are many different possible interpretations or semantic content for this "surveillance value," including: overall uncertainty about the individual's identity; probability that the individual is currently engaged in surveillance-worthy activities; similarity to a large historical database of nefarious actors; and so forth. The autonomous monitoring system may well need to be sensitive to these differences; it could exhibit significant biases—statistical, moral, and legal—if it incorrectly interprets the "surveillance value" module output.

3 Responses to Algorithmic Bias

Algorithms in autonomous systems present many "surfaces" through which different types of bias may enter the system. Because algorithmic bias is not a single monolithic thing, we must be careful about making unqualified assertions of bias, or even more colloquial appeals to notions of neutrality and objectivity. Instead, claims of bias—particularly claims of negative or pernicious bias—require concrete specifications of the relevant standard(s) or norm(s), as well as consideration of the source(s) of bias. In addition, we must also consider the role of the algorithm in question, and its outputs, within the overall system. In particular, there are cases in which algorithmic bias on one dimension can contribute to appropriate performance on a more important dimension, or where the bias or deviation is important in enabling the overall system to achieve desired goals in ways that conform to relevant ethical and legal standards.

3.1 Identifying Problematic Bias

The first step in potential mitigation efforts is to assess whether a given bias is even problematic when all things are considered. As we have previously seen, there are cases in which, say, a degree of statistical algorithmic bias might be necessary in order to reduce or eliminate moral algorithmic bias. And if this statistical bias is relatively innocuous or minor, then we might well judge that it poses no problem (for us, given our goals). That is, we might have statistical

algorithmic bias that is neutral in impact on our values, and that provides some other significant benefit. In fact, we might even actively work to create a statistical algorithmic bias, as noted earlier in the context of a decision system that influences the future population. In that case, the statistical algorithmic bias enables us to reduce a moral *societal* bias. These are concrete instances in which algorithmic bias may be socially or ethically desirable, and so the aspirational goal of algorithmic neutrality would actually be detrimental once all things are considered.

Even if all algorithmic biases somehow could be eliminated, we should not assume that doing so would be beneficial or desirable. The complexity of these cases argues for caution when considering whether and how to approach mitigation in particular cases. These considerations will typically be very challenging and complex, as they require us to consider the relative values that we assign to different aspects of a problem. They are also complicated by the fact that diverse societies exhibit significant variation in both immediate and higher-order relevant values.

As an example of this complexity, consider the seemingly straightforward question of what moral standard should be used when autonomous vehicles face life-and-death decisions about how to distribute risk over the vehicle's passengers and people outside of the car. There is clearly diversity and argument about the right answer to this question, as seen in the large literature on the so-called "Trolley Problem." Some argue that the best ethical standard would treat all lives equally, and thereby require the vehicle to attempt to minimize the number of casualties regardless of where they are located. Others argue that special weight can be given to the vehicle's passengers, and so the vehicle can choose actions that are most likely to minimize harm to them. Any design choice will thereby lead to a system that someone thinks is biased. For example, autonomous vehicles that prioritise passenger safety would exhibit a moral bias according to the standards of the former position, while proponents of the second view would regard this behavior as unbiased or appropriate.

Additional considerations may also be relevant to the "all things considered" determination of whether an ethical bias in favor of passengers should be eliminated or corrected. For example, suppose an ethical bias in favor of passengers turned out (after empirical investigation) to be the only way to secure trust in autonomous vehicles. Since this technology is reasonably expected to reduce the current high levels of traffic-related fatalities, then there may be an "all things considered" ethical obligation to act so as to increase the likelihood of their adoption, even if that means using a "local" ethical bias.

These questions are all fundamentally about our values, both individual and societal, and as such, cannot be entirely answered using technology. They are ethical questions in the sense that they concern our human values and goals, and the means that are permissible to use in pursuing them. Nonetheless, they are questions that cannot be avoided as we assess the performance of algorithms and the systems of which they are a part.

In light of this insight, the foundation of any process for assessing the potential for problematic bias must be a robust and comprehensive understanding of the role that an autonomous system is likely to play in the social contexts in which it is deployed, as well as the basic ethical and legal norms that are relevant to that context. Failure to appreciate the full range of relevant ethical and legal norms in force in a context increases the likelihood that autonomous systems will suffer from transfer context or interpretation bias.

3.2 Intervening on Problematic Bias

Although some algorithmic biases are neutral or even desirable, many are problematic and should be mitigated. As noted above, a prerequisite for this mitigation is an understanding of the relationship between the autonomous system and the ethical and legal norms in force in the relevant contexts. If we have such understanding, then we could potentially mitigate simply by improving this relationship. For example, we might restrict the scope of operation for the system in question so that there is no longer a mismatch in system performance and task demands. Or we might attempt to redesign the system to ensure that it operates in better conformity with relevant norms and constraints.

At a more general level, if we determine that some form of bias requires mitigation or response then we have to be willing to consider responses of various kinds. In the best-case scenario, we might develop a novel algorithm that does not exhibit the problematic bias. Or we might be able to use one type of algorithmic bias to compensate for some other ineliminable bias. For example, suppose our measurement or sampling processes in some domain produce an ineliminable training data bias. If we know the nature of this training data bias, then we can use a bias in the algorithmic processing to offset or correct for the data bias, thereby yielding an overall unbiased system. That is, we can try to develop a system that is overall statistically unbiased, even though different components each exhibit algorithmic bias.

In other cases, this kind of balancing or compensation will require adjustments in the system, whether human or machine, that uses or contains the algorithm. For example, consider a case in which algorithmic focus bias leads to a deviation from a moral standard. That is, the algorithm deviates from our ethical norms about what information should be used. Moreover, suppose that this algorithmic bias cannot be eliminated for some reason. In that case, though, an autonomous system or human using the algorithm output could deliberately employ a compensatory bias based in interpretation bias; for instance, the autonomous system might not take action solely on the basis of the algorithm output. More generally, there are multiple ways to combat algorithmic bias when we judge that a response is required. We are not limited only to technological responses.

Of course, we might be in a situation where there simply is no technological, psychological, or social way to fully correct for the problematic sources, features, and biases. Barocas and Selbst [2016] make this point quite vividly in the domain of employment discrimination. Instead, we must

decide between algorithms that exhibit different biases to determine which is the least bad. In many cases, this choice will be between algorithms that are unbiased relative to (a) statistical or performance standards; or (b) moral or legal norms. These choices require that we look outside of the local technology, or even the local users. These decisions require judgments about relative values, and which we think are more important in this context.

Particularly salient examples of this type of choice arise when we have "sensitive" variables (whether morally or legally) that carry statistical information relevant to solving the problem at hand. That is, some variables ought (in a moral or legal sense) not carry information, but they nonetheless do so statistically. In these cases, we must directly confront questions of value, as we cannot achieve an overall algorithm or system that is unbiased relative to every standard: using the variables will violate a moral or legal norm; not using the variables will lead to statistical deviations. In fact, in these cases, it is no longer clear what is meant by the term "unbiased", as that term suggests that we should strive towards an end-state that is not achievable in this situation.

These choices can become even more complex, as the "sensitive" variable might be capable of serving as an informational proxy for a morally unproblematic, though hard to measure, variable or feature [Pedreschi *et al*., 2008]. For example, we typically think that gender is not morally relevant for job performance evaluations, and so a prediction algorithm that includes gender would exhibit a moral bias due to algorithmic focus bias. However, suppose that gender is correlated with some trait *T* that (i) carries information about job performance; (ii) can be used without moral objection; but (iii) is hard to measure or observe. The first two aspects imply that inclusion of *T* in our algorithm would be unproblematic, and probably desirable. The third aspect, however, means that we cannot actually incorporate *T* in practice. Are we instead permitted to use gender in our prediction algorithm? On the one hand, it seems that we still have the moral bias due to algorithmic focus bias. On the other hand, we are using gender only as a proxy variable, so the moral standard is less clear. We do not claim to have a fixed or definite answer to this question; rather, we raise it simply to point out the complexities that can arise even after we answer the exceptionally difficult questions about whether the bias should be minimized or mitigated.

4 Conclusions

Both popular and academic articles invariably present algorithmic bias as something bad that should be avoided. We have tried to show that the situation is significantly more complex than this. There are many different types of algorithmic bias that arise relative to multiple classes of standards or norms, and from many different sources. Moreover, many of these biases are neutral or can even be beneficial in our efforts to achieve our diverse goals. The bare accusation that some algorithm is biased is therefore uninformative, as it tells us nothing about the nature, scope, source, or size of the deviation from one or more norms; whether those norms are statistical, moral, legal, or other; and whether deviating from that standard is objectionable once all things are considered.

In this paper, we have developed a taxonomy of different kinds and sources of algorithmic bias in an attempt to isolate possible reasons or causes. The different sources are clearly not mutually exclusive, nor do we claim exhaustivity for our taxonomy, though we believe that it covers the vast majority of cases of algorithmic bias. Importantly, algorithmic bias can arise at every stage of the development-implementation-application process, from data sampling to measurement to algorithm design to algorithmic processing to application in the world to interpretation by a human end-user or some other autonomous system. And each entry point for algorithmic bias presents a different set of considerations and possibilities.

As we saw throughout this paper, the taxonomy is not a mere labeling exercise, but rather provides guidance about ways to mitigate various forms of algorithmic bias when they arise. For example, we can often compensate for algorithmic bias in one stage with algorithmic bias in a different one. More generally, we need to think about algorithmic bias (with respect to various norms) in terms of the whole system, including the consumer—human or machine—of the algorithm output. The "ecosystem" around an algorithm contains many opportunities for both the introduction of bias, and also the injection of compensatory biases to minimize the harms (if any) done by the algorithmic bias.

Acknowledgments

Thanks to the anonymous reviewers from the IJCAI Special Track on AI & Autonomy for their valuable comments on an earlier version of this paper.

References

[Arkin *et al*., 2012] Ronald C. Arkin, Patrick Ulam, and Alan R. Wagner. Moral decision-making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE*, 100(3): 571-589. March 2012.

[Barocas and Selbst, 2016] Solon Barocas and Andrew D. Selbst. Big Data's disparate impact. *California Law Review*, 104: 671-732. 2016.

[Garcia, 2016] Megan Garcia. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4): 111-117. Winter 2016.

[Geman *et al*., 1992] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4: 1-58. 1992.

[Kirkpatrick, 2016] Keith Kirkpatrick. Battling algorithmic bias. *Communications of the ACM*, 59(10): 16-17. October 2016.

In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017) forthcoming*.

[Pedreschi *et al*., 2008] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. *Proceedings of KDD 2008*.

[ProPublica, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing