Synthese (2021) 198:9941–9961
https://doi.org/10.1007/s11229-020-02696-y

Algorithmic bias: on the implicit biases of social technology

Gabbrielle M. Johnson[1]

Received: 18 January 2020 / Accepted: 11 May 2020 / Published online: 20 June 2020

Abstract
Often machine learning programs inherit social patterns reflected in their training data without any directed effort by programmers to include such biases. Computer scientists call this *algorithmic bias*. This paper explores the relationship between machine bias and human cognitive bias. In it, I argue similarities between algorithmic and cognitive biases indicate a disconcerting sense in which sources of bias emerge out of seemingly innocuous patterns of information processing. The emergent nature of this bias obscures the existence of the bias itself, making it difficult to identify, mitigate, or evaluate using standard resources in epistemology and ethics. I demonstrate these points in the case of mitigation techniques by presenting what I call 'the Proxy Problem'. One reason biases resist revision is that they rely on proxy attributes, seemingly innocuous attributes that correlate with socially-sensitive attributes, serving as proxies for the socially-sensitive attributes themselves. I argue that in both human and algorithmic domains, this problem presents a common dilemma for mitigation: attempts to discourage reliance on proxy attributes risk a tradeoff with judgement accuracy. This problem, I contend, admits of no purely algorithmic solution.

Keywords Bias · Algorithmic bias · Social bias · Machine bias · Implicit bias

1 Introduction

On March 23rd, 2016, Microsoft Corporation released Tay, an artificial intelligence (AI) Twitter chatbot intended to mimic the language patterns of a 19-year-old American girl. Tay operated by learning from human Twitter users with whom it interacted. Only 16 hours after its launch, Tay was shut down for authoring a number of tweets endorsing Nazi ideology and harassing other Twitter users. Among the inflammatory tweets were those saying "Hitler was right", those endorsing then-Republican-nominee Donald Trump's proposal that "we're going to build a wall", various derogatory remarks

✉ Gabbrielle M. Johnson
gmjohnson@nyu.edu

[1] New York University, New York, USA

about feminists, as well as claims that "9/11 was an inside job". When asked about how Tay developed such a noxious personality, Microsoft responded, "as [Tay] learns, some of its responses are inappropriate and indicative of the types of interactions some people are having with it".[1] In other words, Tay's personality was inherited from the individuals with whom it was engaging.

The story of Tay highlights an obstacle facing developers of machine learning programs: *algorithmic bias*. An AI like Tay, which uses machine learning to capitalize on (or "learn" from) statistical regularities in human-generated datasets, tends to pick up social patterns that manifest in human behavior and that are reflected in the data on which it is trained. In many of these cases, we have reason to suspect that programmers are not explicitly writing biases toward marginalized demographics into their software's code.[2] Instead, it appears the biases in some sense *implicitly emerge* from the algorithms' operating on the data, mimicking the biases reflected in the data themselves. The existence of algorithmic biases undermines the assumptions that computer-based decision-making is more objective and accurate than human decision-making or that it is entirely free from the social and political values humans exhibit. The purpose of this paper is to explore further the connections between the algorithmic and human domains, and to argue that in etiology, operation, evaluation, and mitigation, human biases and algorithmic biases share important similarities. Understanding these similarities, I contend, is critical to succeeding in philosophical and practical projects concerning the amileoration of bias in either domain.

A core similarity of algorithmic biases and human cognitive biases is that both kinds of bias can emerge out of seemingly innocuous patterns of information processing. Investigations into the nature of these emergent biases in the domain of machine learning reveal at least two obstacles for computer programmers: first, often machine learning programs instantiate so-called "black box" algorithms, i.e., those where it is difficult (and arguably impossible) for human observers to describe the rationale for a particular outcome; second, algorithmic biases rely on *proxy attributes*: seemingly innocuous attributes that correlate with socially-sensitive attributes, serving as proxies for the socially-sensitive attributes themselves. These two obstacles exist also in the domain of human cognition, obscuring the existence of biases and making it difficult to identify, mitigate, or evaluate biased decision-making using standard resources in epistemology and ethics. In this paper, I focus especially on the second obstacle, and I argue that the flexibility in the representational states and processes that constitute a bias's operation give rise to what I call 'the Proxy Problem'. Biases that operate on proxies cannot be mitigated using any overt filtering techniques, since eliminating any explicit references to, e.g., race, will be ineffective, as the decision-making procedure can simply substitute a proxy, e.g., zip codes, resulting in similar discriminatory effects. In order to eliminate biases of this sort, I argue we need to focus not on revising particular representational states, but instead on breaking down reliable patterns that exist in the interactions between innocuous representational states and the societal patterns that they encode. However, from this, a second and more pro-

[1] Price (2016).

[2] See, for example, O'Neil (2016, p. 154)'s discussion of discriminatory errors in Google's automatic photo-tagging service.

found dilemma arises for both algorithmic and human decision-making: attempts to discourage reliance on proxy attributes risk a tradeoff with judgement accuracy.

The paper proceeds as follows. In Sect. 2, I present cases where algorithmic biases mimic patterns of human implicit bias. In Sect. 3, I explain how algorithmic biases can manifest without a program's containing explicit stereotype rules, which I demonstrate using a simple toy model of the *k*-nearest neighbors (kNN) learning algorithm. Then, I argue that the same applies to human implicit bias, i.e., some cognitive biases influence an individual's beliefs about and actions toward other people, but are nevertheless nowhere represented in that individual's cognitive repertoire. These similarities highlight the flexibility in the states and processes that give rise to bias in general. In Sect. 4, I demonstrate how this flexibility in structure causes the Proxy Problem. Finally, I argue that to resolve this problem, one cannot merely focus on revising the inner-workings of the decision-making procedures. Instead, I argue, we must refocus mitigation techniques to break down the reliable patterns of interaction among states within these processes and the wider environment in which they're embedded, a point I demonstrate by showing how one might frustrate the reliance on proxy attributes like skin color and dressing feminine in making decisions about whether someone is a worthy presidential candidate. I end by considering the aforementioned tradeoffs of this sort of approach in real-world decision-making.

This paper takes place against the backdrop of a theory I develop in other work, wherein biases of many varieties form a natural kind. According to this theory, bias exists everywhere induction does, and thus, the world of computational bias is vast: biases can be cognitive, algorithmic, social, non-social, epistemically reliable, epistemically unreliable, morally reprehensible, or morally innocuous. I contend that understanding the features common to this natural kind *bias*, which algorithmic bias shares, is critical to making progress on more practical aims that engagement with biases demands. If ultimately our aim is to eliminate biases that are problematic—either epistemically or morally—then our first step should be to understand the origins and operation of biases more generally. Recognizing this natural kind helps to highlight that many of the roadblocks facing the amelioration of harmful biases are not unique to any particular kind of bias, algorithmic included. Indeed, as will become evident by the end of the paper, there are no purely algorithmic solutions to the problems that face algorithmic bias.

2 Evidence for algorithmic bias

I begin by surveying some of the evidence that suggests machine biases mimic typical bias patterns formed from human implicit biases, laying the groundwork for commonalities between the two.

Consider a study by Caliskan et al. (2017) on word-embedding machine learning. This study found that parsing software trained on a dataset called "the common crawl"—an assemblage of 840 billion words collected by crawling the internet—resulted in the program producing "human-like semantic biases" that replicated well-known trends in results of indirect measures for human implicit biases. These biases included the tendency to more often pair stereotypical female names with fam-

ily words than career terms, stereotypical African-American names with unpleasant words rather than pleasant words, and stereotypical male names with science and math terms rather than art terms. A plausible explanation of this phenomenon comes from looking at the patterns within the training data themselves, i.e., patterns in online language use. For example, computer scientist Seth Stephens-Davidowitz's analyses of Google data trends show people are two-and-a-half times more likely to search "Is my son gifted?" than "Is my daughter gifted?". This suggests that online texts encode human social biases that often associate males with inherent intelligence.[3] It seems that the word-embedding software picked up on and mimicked these patterns. On this result, co-author Arvind Narayanan writes, "natural language necessarily contains human biases, and the paradigm of training machine learning on language corpora means that AI will inevitably imbibe these biases as well".[4]

Examples of similar phenomena abound. In 1988, St George's Hospital Medical School's Commission for Racial Equality found a computer program used for initial screenings of applicants "written after careful analysis of the way in which the staff were making these choices" unfairly rejected women and individuals with non-European sounding names.[5] Similarly, a study by Datta et al. (2015, p. 105) found that Google's ad-targeting software resulted in "males [being] shown ads encouraging the seeking of coaching services for high paying jobs more than females." And finally, a study by Klare et al. (2012) demonstrated that face recognition software, some of which is used by law enforcement agencies across the US, use algorithms that are consistently less accurate on women, African Americans, and younger people.[6] As the use of machine learning programs proliferates, the social consequences of their biases become increasingly threatening.[7] One particularly jarring example of these consequences was highlighted by a 2016 ProPublica analysis that revealed software being utilized across the country to predict recidivism is biased against black people, often associating African American attributes with crime, a common result found in implicit bias measures.[8] The algorithm under investigation was nearly twice as likely to falsely flag black defendants as future criminals than white defendants, while inaccurately labeling white defendants as low-risk more often than black defendants.[9]

[3] Stephens-Davidowitz (2014).

[4] Narayanan (2016).

[5] Lowry and Macpherson (1988), p. 657.

[6] See also Wu and Zhang (2016).

[7] For a comprehensive overview of the current state of affairs regarding machine learning programs in social technology, see O'Neil (2016).

[8] See, for example, Eberhardt et al. (2004).

[9] Angwin et al. (2016). The exposé concerned the computer software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). COMPAS has received an enormous amount of attention in philosophy, machine learning, and everyday discussions of algorithmic bias, much of which is beyond the scope of this paper. Important for my purposes, however, is the fact that many of the issues it raises are not unique either to it or to algorithmic decision-making in general. For example, recent work by Kleinberg et al. (2016) and Chouldechova (2016) identifies the three intuitive conditions any risk-assessment program must achieve in order to be fair and unbiased. These criteria include first, that the algorithm is *well-calibrated*, i.e., if it identifies a set of people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances; second, that it *balance the positive class*, i.e., the average score received by people constituting positive instances should be the same in each group;

These examples demonstrate that machine bias exists and that the patterns of such bias mimic well-known implicit bias patterns in humans. However, we need not from the existence of these biases infer that programmers are writing explicitly racist or sexist code. Instead, it is possible that such biases emerge out of the operation of seemingly innocuous code paired with statistical regularities of the training data. These cases of algorithmic bias can be demonstrated with a toy model using a simple k-nearest neighbors algorithm, which I turn to next.

3 Varieties of bias

3.1 Algorithmic bias

I now present a simple walkthrough of how machine learning programs operate.[10] Machine learning programs come in three basic forms: supervised learning, unsupervised learning, and reinforcement learning. In what follows, I begin by focusing on the simpler cases of supervised learning programs. I return to the more complicated cases at the end of the section. There are two main stages of a supervised learning program's operation: first a training phase, followed by a test phase. During the training phase, the program is trained on pre-labeled data. This affords the program the opportunity to "learn" the relationships between features and labels. The second phase consists in applying the resulting predictive model to test data, which outputs a classification for each new test datum on the basis of its features.

For example, imagine that you're creating a program that you intend to use for the simple classification task of predicting whether an object is a ski or a snowboard. You might begin by identifying features of skis and snowboards that you think are relevant for determining an object's classification into either category. In theory, you can choose an indefinite number of relevant features.[11] For our purposes, we'll focus on just two:

Footnote 9 continued
and third, that it *balance the negative class*, i.e., the average score received by people constituting the negative instances should be the same in each group (Kleinberg et al. 2016, p. 2). Strikingly, Kleinberg et al. (2016) demonstrate that in cases where base rates differ and our programs are not perfect predictors—which subsumes most cases—these three conditions necessarily trade off from one another. This means most (if not all) programs used in real-world scenarios will fail to satisfy all three fairness conditions. There is no such thing as an unbiased program in this sense. More strikingly, researchers take this so-called 'impossibility result' to generalize to all predictors, whether they be algorithmic or human decision-makers (Kleinberg et al. 2016, p. 6; Miconi 2017, p. 4). For a compelling argument of how best to prioritize different notions of fairness, see Hellman (2019).

[10] Ideally, I would present a walkthrough of the operation of one of the algorithms discussed above. Unfortunately, providing a detailed analysis of one of these algorithms is difficult, if not impossible, since information relating to the operation of commercial machine learning programs is often intentionally inaccessible for the sake of competitive commercial advantage or client security purposes. Even if these algorithms were made available for public scrutiny, many likely instantiate so-called 'black-box' algorithms, i.e., those where it is difficult if not impossible for human programmers to understand or explain why any particular outcome occurs. This lack of transparency with respect to their operation creates a myriad of concerning questions about fairness, objectivity, and accuracy. However, these issues are beyond the scope of this discussion.

[11] The problem of which features are most relevant for some classification task will itself raise important philosophical issues, some of which I return to in my discussion of the Proxy Problem in Sect. 4.

length and width.[12] We begin stage one by training our program on pre-labeled data called 'training data'. These include many instances of already-categorized objects. We can represent the relationships between the relevant features and classifications for the known data by plotting them on a two-dimensional feature space, as shown in Fig. 1.

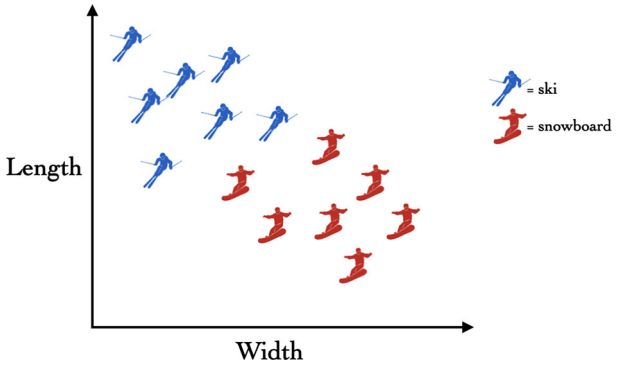

Fig. 1 Training phase—skis versus snowboards

Each data-point specifies two things: the *feature values*, i.e., values corresponding to its length and width; and a *class label*, i.e., a label corresponding to its classification as a ski or snowboard. Per Fig. 1, the objects in the test data that are longer but not as wide tend to be skis, while the objects that are wider but not as long tend to be snowboards.

In the next phase of the program, the algorithm gets applied to new, unclassified data called 'test data', and the aim of the program is to classify each datum as either a ski or a snowboard on the basis of its feature values. One way for the program to do this is to classify new instances on the basis of their proximity in the feature space to known classifications. For example, say we had a new object that we didn't know was a ski or a snowboard, but we did know had a certain length and width. Based on these feature values, the datum is plotted in the figure space as shown in Fig. 2.

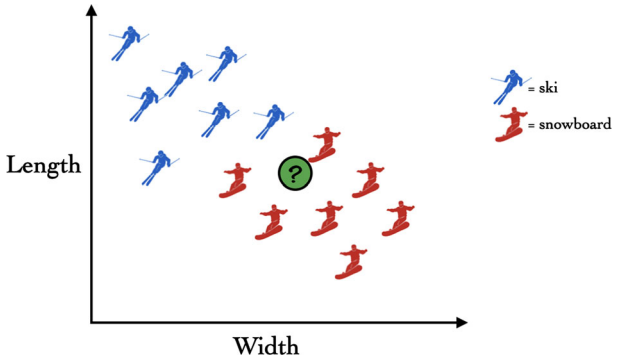

Fig. 2 Test phase–skis versus snowboards

[12] This case is discussed in more detail by Daum III (2015, pp. 30–32).

The program can then use the test datum's relationship to the other data points to make a novel classification. There are a variety of ways this new data point is related to the other data points, and thus a variety of methods on which to make the new classification. One intuitive method is to identify a number k of its nearest neighbors and allow them to "vote". For example, if we set k to 5, then the program will decide on the basis of the five nearest neighbors in the feature space. As evident in Fig. 2, the five nearest neighbors comprise one ski and four snowboards. Since the majority vote in this case results in snowboards, the program classifies the test instance as a snowboard.

This same method can be used for any number of classification tasks based on relevant attributes, including those overlapping with stereotypical judgements based on members of marginalized social groups. For example, imagine an engineer is creating a program that classifies individuals as good or bad with computers. Let's say she thinks one relevant property for determining this classification is a person's age. Thus, she trains the program on many instances of individuals she labels as a certain age and as either good or bad with computers. Figure 3 shows how these instances might end up plotted on a one-dimensional feature space.

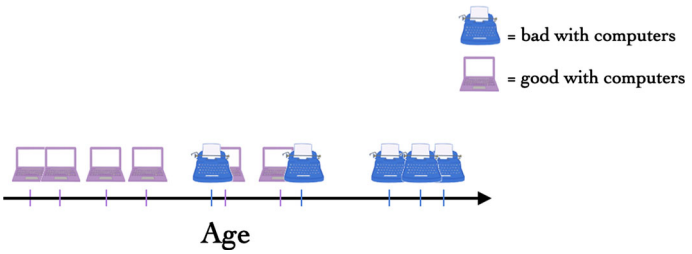

Fig. 3 Training phase—good versus bad with computers

Per Fig. 3, the data in this example are skewed: those individuals who are bad with computers are clustered near the end of the line, while those good with computers are near the beginning. Prima facie, this data pattern appears problematic in some sense. There are many ways test data can come to be patterned like this, and it is worth taking time to explore a few of them. Two ways the data might come to be patterned like this involve bad data collection practices. For example, imagine that the programmer pulled training data from a local library on the same day the library was hosting an after-school event during which high-schoolers provide social media training to residents of a local assisted living facility. In this case, we would not expect the data to be representative of the general population: her sample would have a greater proportion of both tech-savvy young people (compared to the general population of young people) and elderly individuals who struggle with computer technology (compared to the general population of elderly individuals). The second way our training data can be skewed via problematic data collection practices is by a failure to accurately label the samples. For example, if the person labeling the training data in our case was—due to her own biases and prejudices—more likely to label elderly individuals as bad with computers even when they weren't, then we should still expect a mismatch between

the training data and the real world. Although seemingly contrived, these examples demonstrate an important lesson: a machine learning program is only as good as the data on which it is trained, giving rise to the oft-cited motto "garbage in, garbage out". If the data going into the training period are problematic, then we can expect the generalizations the program makes based on those data to be problematic as well.

However, there is a third possibility for how data can come to be patterned like this that does not involve bad data collection practices, but is of great philosophical interest nonetheless. It is arguably the source of many of the real-world cases of algorithmic bias discussed in this paper. It occurs when even randomly selected data might reflect social biases because such patterns are ubiquitous in the environment. For example, if it turns out that, in general, the elderly are statistically more likely to be bad with computers, then even carefully collected data might reflect this problematic trend. Crucially, this claim itself makes no commitments about the source of the problematic trend in the wider environment, and certainly need not entail any claims about the natural dispositions of elderly individuals. Instead, the trend that elderly people are more likely to be bad with computers—and other objectionable patterns concerning vulnerable demographics—might be a result of structural and societal discrimination. Indeed, given that our environment is widely shaped by historical patterns of injustice and discrimination, we can expect many problematic social patterns to be ubiquitous in this way. Thus, one of the primary questions for philosophers of machine learning and computer programmers to address is the following: particularly in the social domains in which this technology is employed, does there exist such a thing as non-problematic training data and, if so, how do we attain such data? If not, there seems to be a philosophically robust sense in which, at least in these domains, there is no alternative to "garbage in", and, thus, there will likewise be no alternative to "garbage out".

We see this lesson of "garbage in, garbage out" unfold when we again apply our kNN algorithm to this example, as shown in Fig. 4. In this case, the five nearest neighbors to the target will include a majority of individuals who are bad with computers. Thus, the new individual will, on the basis of their age, be labeled as bad with computers.

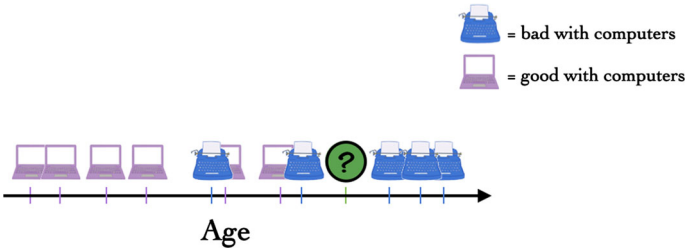

Fig. 4 Test phase—good versus bad with computers

With those individuals who are older patterning roughly with those who are bad with computers (and vice versa for young individuals and good with computers), we could think of this entire dataset as reflecting the social generalization (or, loosely, the "stereotype") that elderly people are bad with computers. Importantly though, notice that this content is never explicitly represented as a rule in the program's code.

Instead, the algorithm operates on a simple Euclidean distance calculation involving the proximity of the test instance to the training instances in the feature space. Rather than the stereotype being explicitly represented as a rule, it *implicitly emerges* out of the distribution of training instances in the feature space. It is this distribution of training instances that gives rise to a bias toward the elderly.

Most important, in cases where the stereotype rule isn't explicitly represented in a program, the program still operates *as if* it represented such a rule: it still classifies individuals as being bad with computers on the basis of their age, with elderly individuals being more likely to be labeled as bad with computers. I call biases that operate in this way *truly implicit biases*, and algorithmic biases—such as those arising in the examples above—are one form such biases can take.[13]

In what follows, I draw an analogy of this case to human cognitive biases to demonstrate that, as with algorithmic bias, cognitive biases may be truly implicit. This flexibility in the states and processes that give rise to bias creates complications in the identification, evaluation, and mitigation of bias in general.

3.2 Cognitive bias

Although occurring in different domains, algorithmic biases and human biases have important structural commonalities. To demonstrate this, imagine a paradigmatic case of social cognitive bias. For example, imagine that a fellow academic attempts to help his elderly colleague Jan join a Zoom interview.[14] When asked why he did this, he explains that it is because Jan is elderly, and he believes elderly people are bad with computers. Here, the inference the colleague is making is straightforward:

(*i*) Jan is elderly.
(*ii*) Elderly people are bad with computers.
∴ (*iii*) Jan is bad with computers.

This case is typical of cases of so-called 'explicit bias', since the colleague is completely aware that he is drawing conclusions about Jan based on his beliefs about the elderly and Jan's belonging to that group. The notion 'bias' is often used to refer to different aspects of decision-making indicative of bias.[15] For simplicity, my focus will be on the collection of states and processes that, when coupled with a belief that some individual belongs to a social group, produces a conclusion that that individual has properties stereotypical of the social group. I call this collection of states and processes 'the bias-construct', and the bias-construct in this example is the stereotype belief that elderly people are bad with computers (together with whatever inferential processes are necessary to derive the conclusion).

Largely, extant theories of social bias regard *implicit* biases as involving bias-constructs that we are not aware or conscious of and *explicit* biases as involving those that we are. This is a claim that implicit and explicit biases differ with respect to

[13] See Johnson (2020).

[14] This example and parts of its description are borrowed from Johnson (2020).

[15] See Holroyd and Sweetman (2016) for discussion and examples.

their conscious accessibility.[16] Bias-constructs that are unconscious might be, on the one hand, stored and represented like other mental states; or, more curiously, they might be "merely encoded" in the patterns of other states being processed. In this case, the states and processes underlying them are analogous with respect to their representational status to algorithmic biases—they are *truly implicit*.[17]

To see this point, it helps to notice that distinct bias-constructs can systematically relate bias-inputs and bias-outputs in similar ways. Imagine another colleague also attempts to help Jan join a Zoom interview, but when asked why she did this, she is unable to offer any explanation.[18] Like the previous colleague, she too thinks that Jan is elderly and comes to the conclusion that Jan needs help with the Zoom interview; however, unlike the previous colleague, she reports believing that elderly individuals are just as good with computers as anyone else is. Here, the inference this colleague is making is less clear:

(i) Jan is elderly.
(iii) Jan is bad with computers.

One possible explanation for the transition between (i) and (iii) is the existence of an unconscious stereotype belief that elderly people are bad with computers. However, crucially, other combinations of states and processes can just as easily fill this role. For example, it might be that what is happening in this colleague's head is a combination of states and processes that are similar to the kNN algorithm discussed above that classified individuals as being bad with computers on the basis of their age and similarities to past individuals this subject has encountered. If so, then as in the case of algorithmic bias, the transition from (i) to (iii) could occur without ever explicitly representing a stereotype. That is, principles governing the transitions from the belief that Jan is elderly to the belief that she's bad with computers are plausibly merely encoded in just the same way as the algorithmic biases above. In this case, her bias is truly implicit.[19]

There are several important takeaways from the discussion in this section. The first concerns the generality of bias. The features that are common to cases of problematic social biases generalize to non-problematic, non-social biases. Notice that the features that gave rise to a bias in the example of elderly people being bad with computer and young people being good with computers were present also in the example of skis being long and thin and snowboards being short and wide. In both cases, generalizations emerged not from an explicitly represented stereotype-like rule, but rather from run-of-the-mill operations on datasets that encoded the relevant stereotype-like patterns. As

[16] Recently theorists including Gawronski et al. (2006) and Hahn et al. (2014) have disputed that implicit biases are in fact unconscious. I don't take up the dispute in detail here because it is largely irrelevant for my claims regarding the *representational nature* of bias-constructs.

[17] Corneille and Hutter (2020) survey conceptual ambiguities in the use of 'implicit' in attitude research. According to them, among the possible interpretations present in standard literature are implicit as automatic, implicit as indirect, and implicit as associative (as well as hybrids). The interpretation of implicit as truly implicit, i.e., non-representational, is largely ignored or altogether absent from these interpretations. See also Holroyd et al. (2017), pp. 3–7.

[18] More standard in cases like this, she might instead offer an explanation that, upon investigation, is revealed to be confabulation.

[19] See Johnson (2020).

demonstrated, these same features of algorithmic bias generalize to non-algorithmic, cognitive biases as well.

These points evoke an understanding of a natural kind *bias*, under which problematic social algorithmic and cognitive bias emerge as species. This broader kind is normatively neutral and explanatorily robust. Against common usage, it includes biases that are epistemically reliable and morally unproblematic. This is in tension with the notion of bias employed in many everyday discussions of algorithmic bias that bakes into the term negative connotations. Understandably, some theorists might prefer to retain this negative valence associated with the notion *bias*. However, I believe that resisting this tendency so as to see the commonalities of the general kind *bias* shepherds in a variety of theoretical upshots, which I regard as critical to achieving the common aim of combating morally or epistemically problematic biases. On this general understanding, biases are necessary solutions to underdetermination, and thus, bias exists anywhere induction does. The overarching reason for adopting a normatively neutral notion of bias is that regarding bias as constitutively problematic implies that the alternative is to adopt a form of objectivity that would make inductive reasoning impossible.[20]

Induction in the face of underdetermination arises in a variety of domains, and studying how biases are employed as solutions in one can be fruitful for understanding their operation in another. Consider the following helpful analogy between the etiologies of visual perceptual biases and problematic social biases. Imagine a simple robot built to navigate through an obstacle course. A robot trained on machine learning models intended to mimic the human visual perceptual system will face an underdetermination problem, since the evidence of the 3-D physical environment that it has access to will be limited to mere 2-D proximal stimulations of light pattern registrations. Thus, it will necessarily have to adopt various biases about the environment. In our natural environment, for example, light tends to come from above. So, if a robot is going to successfully navigate the physical environment, it will need to pick up on and utilize this assumption when making judgements about its surroundings. This assumption isn't perfect, and it wouldn't work in environments where light doesn't come from above, but in our environment, it tends to get things right. Drawing the analogy to social biases, our social environment is shaped by a variety of racist and discriminatory practices. So, if a machine learning program is aiming to make predictions in line with our current social landscape—i.e., built to navigate our current social environment—it necessarily adopts and utilizes assumptions that mimic patterns presently existing in the data on which it is trained. Thus, assumptions that encode problematic stereotypes will inevitably be adopted and perpetuated by machine learning programs. This will continue until we shape a new social environment for which such assumptions will be ill-placed.[21]

[20] I discuss these points in greater detail in Johnson (2020), following considerations presented by Antony (2001, 2016). For other arguments in favor of and against including normative and accuracy conditions in the definition of bias and the related notion *stereotype*, see Munton (2019b), Beeghly (2015), and Blum (2004).

[21] I return to these points at the end of §4.

Another aspect of the generality of bias is the flexibility in states and processes that give rise to it.[22] I've already demonstrated this flexibility by giving instances where the same bias against the elderly can emerge either as an explicitly represented stereotype rule or as a combination of innocuous rules on a particular data set. This point is compounded as we look at more examples, especially in the algorithmic domain. It is important to note that the toy kNN model I have presented is—for the ease of exposition and understanding—much simpler than real-world machine learning programs. In the range of more complicated cases that make up actual machine learning programs, e.g., those relying on high-dimensional feature spaces that encode collections of a great number of feature values, it seems plausible that these points about flexibility will be even more pressing. For example, deep neural nets and unsupervised learning programs are less likely to be adopting anything close to the explicitly represented stereotype-like rules that are often assumed in discussions of bias. They are also unlikely to use category labels that correspond to the overt target labels that show up in stereotype generalizations. Yet, they exhibit problematic biases all the same. It is this flexibility that will pave the way for the Proxy Problem, which I turn to next.

4 The proxy problem

At this point, I've argued that similar discriminatory patterns can emerge from a range of decision-making procedures with variable states and processes. Here, I discuss a problem that emerges from this fact: the Proxy Problem.

Machine learning programmers have long struggled with eliminating algorithmic biases that are based on so-called 'proxy attributes': seemingly innocuous attributes that correlate with socially sensitive attributes, serving as proxies for the socially-sensitive attributes themselves.[23] Often, engineers attempt to protect disadvantaged social groups by preventing classification algorithms from performing tasks on the basis of socially-sensitive features in cases where using these features would be discriminatory. For example, it would be both legally and morally problematic to allow a program to categorize candidates as either eligible or ineligible for a mortgage loan based on those candidates' races. As a result, programmers attempt to prevent any explicit reliance on race by including filters that block the program from labeling individuals on the basis of race. However, often these filters fail to prevent the program from adopting proxy attributes that correlate with the socially-sensitive attributes. For example, rather than labeling some individual as 'African American', the program might label them based on their zip codes.[24] Since neighborhood demographics are

[22] In the case of implicit human cognitive bias, this point about flexibility is often raised under the heading of the heterogeneity of bias. See, in particular, Holroyd and Sweetman (2016).

[23] See, for example, Adler et al. (2016)'s discussion of how to audit so-called 'black box' algorithms that appear to rely on proxy attributes in lieu of target attributes.

[24] To take an example from earlier, the recidivism algorithm COMPAS produces patterns discriminatory toward African Americans despite race demographic information not being explicitly included in the information given to it about each defendant. Instead, it has access to other features that collectively correlated with race.

often racially homogeneous, a person's zip code often correlates with their race. Thus, an algorithm utilizing the former might operate very similar to it utilizing the latter.[25]

Likewise, a proxy attribute could be used in the bad-with-computers classification algorithm above to produce similar results. Instead of explicitly relying on an individual's age to plot them in the feature space, a programmer might instead rely on a proxy, say their degree of trust in Fox News, as shown in Fig. 5.[26]

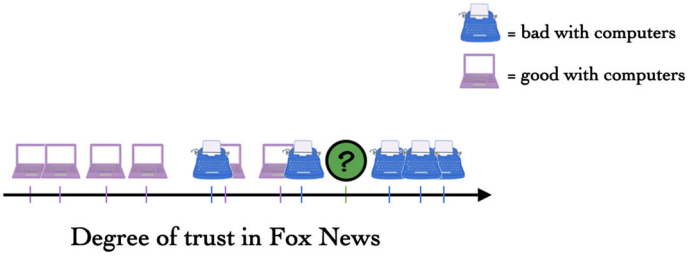

Fig. 5 Proxy—good versus bad with computers

Here, the algorithm need not make any explicit reference to age. Most important, in cases where the socially sensitive attribute isn't explicitly represented in a program, the program again operates *as if* it represented such a content: it still classifies individuals in such a way that elderly individuals are more likely to be labeled as bad with computers.

The same can be said for cognitive biases that rely on proxy attributes. Imagine yet another colleague whose overt interactions with Jan are the same as the previous two, but whose reasoning about her is very different. This colleague doesn't really consider Jan's age at all, but does know that she has a high degree of trust in Fox News. We can imagine his overt reasoning about Jan looks something like this:

(i) Jan really trusts Fox News.
(iii) Jan is bad with computers.

This person might never recognize the connection between the attributes of having a high degree of trust in Fox News and being elderly. He also never endorses a stereotype of the form elderly people are bad with computers. Nevertheless, he has a bias that causes him to treat elderly individuals as being bad with computers.

The insight that biases (cognitive or algorithmic) might operate using proxy attributes has important implications for mitigation techniques in both domains. Consider one more classification task. This time, imagine that the goal is to categorize individuals as presidential or non-presidential and that the chosen relevant features the procedure relies on are skin tone and the degree to which a person dresses in a stereotypically feminine manner—imperfect proxies for race and gender, respectively.

[25] For reasons soon to be discussed, programmers are forced to rely on proxies. Moreover, the notion of proxy discrimination is familiar in discussions of discrimination in ethics and law. See, for example, Alexander (1992), pp. 167–173.

[26] The median age of Fox News viewers is 65 (Epstein 2016). Although age might not correlate exactly with perceived trustworthiness, we can assume for simplicity that it does.

Just as in the training phases above, this procedure utilizes known instances of the target categorization. In this case, let's assume it has access to all individuals who have run for the presidency. Based on historical discriminatory trends, we can imagine the feature space looks very roughly like Fig. 6.[27]

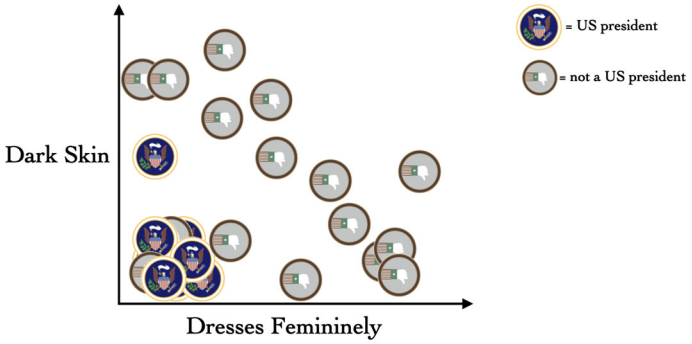

Fig. 6 Training phase—worthy versus not worthy presidential candidate

With these training data, we can predict what is likely to happen in the test phase with an individual who dresses in a stereotypically feminine manner, as shown in Fig. 7.

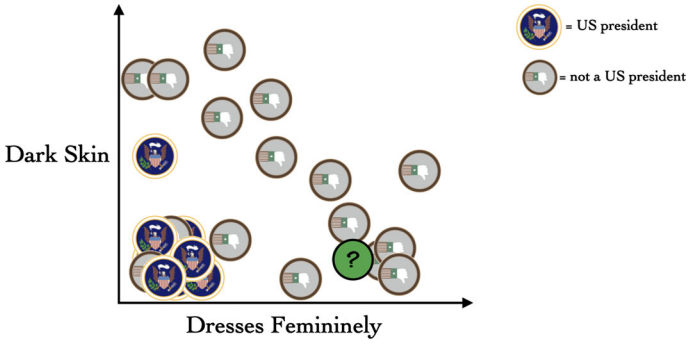

Fig. 7 Testing phase—worthy versus not worthy presidential candidate

Such an individual will, on the basis of their feminine presentation, be categorized as not a U.S. President. Crucially, the proceedure makes no explicit reference to race or gender, but the results of its operation mimic the results of a procedure that does rely on such features.

Although here I use examples that are morally and politically salient or that rely on various intentional choices of individuals, I intend for the points I make to generalize to a variety of other cases. That we use some properties to stand in for others during

[27] This feature space is meant to be a simplification and is not intended to exactly match real-world data.

reasoning is a ubiquitous phenomenon.[28] Often in reasoning we are interested in some deep property that goes beyond the properties we have direct access to. Thus, reliance on proxies will be a necessary feature of most inductive reasoning. In this general sense—where proxy reasoning occurs any time we use some feature to correlate with another—proxy use will likewise be ubiquitous.[29] Notice that this is true also of machine learning programs. Programs built for object recognition, say, are rarely in contact with the actual objects they aim to identify. Instead, they rely on pictures. In this case, some collection of pixel values will act as a proxy for some other feature, perhaps shape, which will then act as a proxy for some target natural kind attribute, being a dog, say. So, proxies are neither rare nor the result of overly intentional decisions; rather, they are often necessary and unwitting.

Computer programmers have long struggled with problematic proxy attributes since programs that rely on them tend to resist any overt filtering techniques. Eliminating any explicit references to a socially sensitive attribute in the program's code will be ineffective, as the program can simply substitute reference to a proxy attribute in its place, resulting in similar discriminatory effects. The program has been tasked with finding the most efficient and reliable way to classify a new object based on the correlation patterns between attributes and target features in the data. Where there are robust correlations between socially sensitive attributes, proxy attributes, and target features and we've ruled out using the socially sensitive attributes, the next best thing for the program to use will be the proxy attributes.

One might think that human implicit biases similarly resist revision. That is, if a mitigation technique operates on overt, socially-sensitive attributes, and the relevant cognitive biases instead rely on proxy attributes, then that mitigation technique will fail. If so, we might borrow mitigation techniques from machine learning. One programming strategy is to curate the training data in such a way that the problematic features bear no straightforward relationship to the relevant categories, making reliance on them for categorization ineffective. If we wanted to likewise frustrate the reliance on the proxies in the example above, one place to start is to introduce more instances of counter-stereotypical exemplars, as shown in Fig. 8.[30]

In this case, relying on whether a candidate dresses in a stereotypically feminine manner will not be a reliable guide in categorizing them as presidential.

[28] Cognitive examples where some attributes can stand in for others (and where this is taken to reflect some quirk underwriting heuristic judgements) are well-studied in empirical work on 'attribute substitution effects'. According to the theory, there are some instances where decisions made on the basis of a target attribute—familiarity, say—are too cognitively demanding. This prompts the cognitive system to shift to a less cognitively demanding task that relies on some irrelevant, but more easily identifiable attribute—e.g., beauty. For discussion, see Kahneman and Frederick (2002) and Monin (2003).

[29] This is a perfectly adequate first-pass analysis of the notion of proxy as it is used in machine learning. See, for example, Eubanks (2018, pp. 143–145, 168)'s discussion of using community calls to a child abuse hotline as a proxy for actual child abuse. However, I believe the key features of proxies that make them a useful concept in our explanations about discrimination go deeper than this first-pass analysis. Ultimately, I take the philosophically rich features of proxy discrimination to stem from an externalism about intentionality and anti-individualism about the natures of representational states, points I discuss further in my "Proxies Aren't Intentional, They're Intentional" (MS).

[30] See Byrd (2019) for a review of the effectiveness of cognitive debiasing strategies that utilize counter-stereotypical training (i.e., counterconditioning) and what this effectiveness entails for theories of social cognitive bias-constructs.

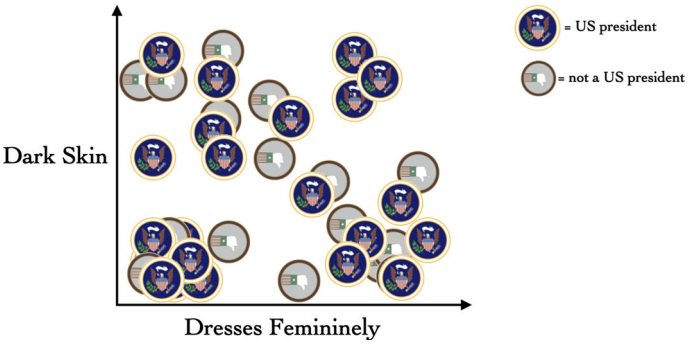

Fig. 8 Counterstereotypical—worthy versus not worthy presidential candidate

This example highlights an avenue for optimism about mitigating truly implicit social biases in cognition and machines. However, there exist several challenges. Firstly, this example appears to naively advocate that we combat biases by changing the overarching social patterns. However, these social patterns are likely themselves the result of the biases we are attempting to ameliorate. So, we can't *simply* advocate that we fix these issues by electing more women and people of color. However, as a more modest upshot, it can serve to bolster a critical insight from equality advocates that *representation matters*.[31]

Secondly, and more problematically, there often exists a complex tradeoff between reliance on proxy attributes and accuracy judgements. Because our world is shaped by historical patterns of oppression, often the best available indicators for some classification task will necessarily serve as proxies for membership in marginalized groups. Take, for example, the fact that women are underrepresented in philosophy. Some hiring committee aware that this is partly the result of explicit and implicit biases toward women in philosophy might anonymize applications in an attempt to eliminate the effects a candidate's gender has on her application. Instead, they decide to rely on "purely objective" features of the application, like publication rate. But these seemingly objective features often correlate with gender, thus making them unwitting proxies. For whatever reasons—whether they be from pervasive individual-level biases or structural injustices—the percentage of published papers by female authors in philosophy is extremely low.[32] So, reliance on publication record will inevitably disadvantage women. But, on the other hand, doesn't publication record seem like a reasonable metric by which to make hiring decisions? If the committee eliminates it

[31] It additionally bolsters the point that combating problematic biases will require a two-pronged solution that focuses both on the individual and structural level. See Antony (2016) and Huebner (2016), as well as the debate between so-called 'structural prioritizers' and 'anti-anti-individualists', including Haslanger (2015, 2016b, a), Ayala Lopez (2016, 2018), Ayala Lopez and Vasilyeva (2015), Ayala Lopez and Beeghly (2020), Madva (2016), and Soon (2019).

[32] Wilhelm et al. (2018) put the percentage between 14 and 16%, despite women constituting closer to 25% of all philosophy faculty and 50% of the general population. See also analyses of the number of publications in philosophy by female authors (and the underrepresentation of women in philosophy in general) presented in Paxton et al. (2012), Saul (2013), Jennings and Schwitzgebel (2017), Nagel et al. (2018), and Kings (2019).

from their decision-making criteria, then they arguably limit their chances of accurately identifying worthwhile candidates, at least insofar as worthwhile is taken to align with traditional metrics of academic success. Moreover, it's not clear that there exist any alternative criteria that wouldn't raise similar concerns.[33]

Thus, a dilemma arises: due to patterns of oppression being so deeply engrained in our social environment, the more we eliminate reliance on proxy attributes in decision-making, the more likely it is that we'll have an inaccurate and ineffective decision-making procedure.[34] This is because our standards for accuracy have been shaped by the social environment, which has itself been shaped by oppressive structures. This tradeoff between demands of epistemic reliability and demands of morality has already received much attention in the domain of human reasoning about members of social groups.[35] It highlights another way that issues surrounding algorithmic bias are not unique to algorithmic decision-making, the extent to which the sources of biases can be common between the two domains, and how theorizing about bias in one domain can enhance our understanding of bias in the other. Future work in the ethics and epistemology of machine and human decision-making procedures that rely on biases will require recognizing this tradeoff and reasoning carefully about how best to balance it. For now, the dilemma makes clear that there are no purely algorithmic solutions to the problems that face algorithmic bias.

5 Conclusion

Inquiries regarding the nature of bias and the utility of comparing its existence in both machine and cognitive domains are still in their infancy. However, preliminary results showing the usefulness of these comparisons have been positive. This paper aimed to highlight just a few of these positive results, while gesturing at the many avenues such comparisons open for further, fruitful philosophical engagement. I end by reflecting on reasons we might be averse to this general comparison, given that it implies humans and computers are capable of similar psychological capacities. For the same reason

[33] I am forced to simplify a complex discussion about what makes for "*accurate* judgements of being a *worthwhile* candidate". For reasons this discussion brings out, traditional metrics of academic success, e.g., standards for tenure promotion, will have baked into them (potentially problematic) preconceptions that will often track structural inequalities. Thus, to be "accurate" with respect to those standards will entail alignment with problematic social trends. We could adopt different measures of success, e.g., contributions to diversity, that don't have baked into them those same preconceptions. However, these other metrics would arguably have baked into them some other preconceptions that are themselves morally- and politically-laden. I discuss in more detail how seemingly neutral notions like 'accuracy' are potentially value-laden in my "Are Algorithms Value-Free?" (MS). Thanks to an anonymous referee for pushing me to be more explicit about these points and for noting how the idea of a "forced choice" between inclusivity and excellence likely requires an overly narrow conception of excellence. See also Stewart and Valian (2018), pp. 212–213.

[34] Indeed, as an anonymous referee points out, this is arguably evidence that a "colorblind" approach that attempts to ignore socially-sensitive features is misguided to begin with. Alternative to this approach could be to explicitly code for the socially-sensitive features, i.e., allow explicit reference to features like race in the program, so as to overtly counter-balance the effects caused by the Proxy Problem. This follows Anderson (2010, pp. 155–156)'s discussion of how color-conscious policies are required in order to end race-based injustice. I agree, but leave more detailed consideration of these alternative strategies for future work.

[35] Gendler (2011), Basu (2019a, b), Bolinger (2018), and Munton (2019a), among others.

we resist saying that a computer has any beliefs, desires, or emotions, we might also want to resist attributing to them biases proper.

Although I think that the similarities I've highlighted facilitate robust predictive and explanatory exchange between biases in the machine and cognitive domains, my account leaves open that there might still be philosophical reasons for distinguishing between them. For example, theories of content in philosophy of language might individuate the representations involved in the input-output profile of human biases from the contents involved in algorithmic biases. Similarly, theories of moral responsibility and blame in ethics and value theory might provide philosophically important reasons for regarding the biases that operate in human agents as entirely distinct from the biases that operate within machine learning programs. There also exist potential empirical discovers that could undermine the comparison, e.g., if it turns out that human cognitive biases are constitutively affective and that artificial intelligence might be incapable of manifesting affective attitudes. My point is not to deny that these other theories might eventually provide reasons for distinguishing between the two, but rather to resist at the onset foreclosing the possibility that these investigations might go the other way, or that there are other explanatory projects for which the identification is apt.

Finally, although I see on the horizon tempting philosophical reasons to avoid saying computers are biased in the same way humans are, I also feel there are pragmatic considerations that motivate this comparison, as it clearly highlights valuable avenues of philosophical inquiry explored in one area that have been unexplored in the other. My discussion in this paper of emergent biases and the Proxy Problem are examples of such avenues. Along similar lines, it seems to me there exist interesting questions regarding what constitutes a bias in the first place and whether it is even possible to entirely eliminate biases within a person or program that likewise benefit from exchange between the two domains.

In summary, this account is not intended to argue that algorithmic and cognitive biases are similar in all respects; rather, it is intended as a starting point for comparisons between the two. This starting point might eventually enable us to see important differences between the two cases, or it might enable us to see the similarities; in both cases, however, it will lend to a better understanding of bias.[36]

[36] I have many to thank for the development of this paper over the years. Firstly, thanks to Tatyana Kostochka for drawing the illustrations used to make the figures. For comments and discussion about the paper, thanks to Josh Armstrong, Rima Basu, Erin Beeghly, Renee Bolinger, Michael Brownstein, Elisabeth Camp, David Chalmers, Sam Cumming, Gabe Dupre, Tina Eliassi-Rad, Maegan Fairchild, Branden Fitelson, Daniel Fogal, Deborah Hellman, Pamela Hieronymi, Justin Humphreys, Amber Kavka-Warren, Seth Lazar, Travis LaCroix, Dustin Locke, Alex Madva, Eric Mandelbaum, Annette Martin, Jessie Munton, Eleonore Neufeld, Sara Protasi, Chelsea Rosenthal, Ronni Gura Sadovsky, Ayana Samuel, Susanna Schellenberg, Susanna Siegel, Seana Shiffrin, Joy Shim, and Annette Zimmermann. Previous versions of this paper were presented at and received valuable feedback from the Vancouver Summer Philosophy Conference at the University of British Columbia, Athena in Action at Princeton University, Philosophy of Machine Learning: Knowledge and Causality at University of California, Irvine, and Bias in Context Four at the University of Utah. Finally, I want to acknowledge the helpful suggestions received from Nick Byrd and an anonymous referee at *Synthese*.

References

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016). Auditing black-box models for indirect influence. In *2016 IEEE 16th international conference on data mining (ICDM)* (pp. 1–10). IEEE.

Alexander, L. (1992). What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review*, *141*(1), 149.

Anderson, E. (2010). *The imperative of integration*. Princeton: Princeton University Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. New York: ProPublica.

Antony, L. (2001). Quine as feminist: The radical import of naturalized epistemology. In L. Antony & C. E. Witt (Eds.), *A mind of one's own: Feminist essays on reason and objectivity* (pp. 110–153). Boulder: Westview Press.

Antony, L. (2016). Bias: friend or foe? In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 1, pp. 157–190)., Metaphysics and epistemology Oxford: Oxford University Press.

Ayala Lopez, S. (2016). Comments on Alex Madvas 'A plea for anti-anti-individualism: How oversimple psychology mislead social policy'. In *Ergo symposium*.

Ayala Lopez, S. (2018). A structural explanation of injustice in conversations: It's about norms. *Pacific Philosophical Quarterly*, *99*(4), 726–748.

Ayala Lopez, S., & Beeghly, E. (2020). Explaining injustice: Structural analysis, bias, and individuals. In E. Beeghly & A. Madva (Eds.), *Introduction to implicit bias: Knoweldge, justice, and the social mind*. Abingdon: Routledge.

Ayala Lopez, S., & Vasilyeva, N. (2015). Explaining injustice in speech: Individualistic vs structural explanation. In R. Dale, C. Jennings, P. P. Maglio, T. Matlock, D. C. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 130–135). Austin: Cognitive Science Society.

Basu, R. (2019a). The wrongs of racist beliefs. *Philosophical Studies*, *176*(9), 2497–2515.

Basu, R. (2019b). What we epistemically owe to each other. *Philosophical Studies*, *176*(4), 915–931.

Beeghly, E. (2015). What is a stereotype? What is stereotyping? *Hypatia*, *30*(4), 675–691.

Blum, L. (2004). Stereotypes and stereotyping: A moral analysis. *Philosophical Papers*, *33*(3), 251–289.

Bolinger, R. J. (2018). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, 1–17.

Byrd, N. (2019). What we can (and cant) infer about implicit bias from debiasing experiments. *Synthese*, 1–29.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv preprint arXiv:1610.07524.

Corneille, O. & Hutter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, 108886832091132.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. In *Proceedings on privacy enhancing technologies* (Vol. 2015(1)).

Daum III, H. (2015). A Course in Machine Learning. https://ciml.info/.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876–893.

Epstein, A. (2016). *Fox News's biggest problem isn't the Ailes ouster, it's that it's average viewer is a dinosaur*. New York: Quartz Media.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police and punish the poor*. New York: St. Martin's Press.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are implicit attitudes unconscious? *Consciousness and Cognition*, *15*(3), 485–499.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*(1), 33–63.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392.

Haslanger, S. (2015). Social structure, narrative, and explanation. *Canadian Journal of Philosophy*, *45*(1), 1–15.

Haslanger, S. (2016a). Comments on Alex Madvas 'A plea for anti-anti-individualism: How oversimple psychology mislead social policy'. In *Ergo symposium*.
Haslanger, S. (2016b). What is a (social) structural explanation? *Philosophical Studies*, *173*(1), 113–130.
Hellman, D. (2019). Measuring Algorithmic Fairness. *Virginia Public Law and Legal Theory Research Paper*, *2019*, 39.
Holroyd, J., Scaife, R., & Stafford, T. (2017). What is implicit bias? *Philosophy Compass*, *12*(10), e12437.
Holroyd, J., & Sweetman, J. (2016). The heterogeneity of implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy volume 1: metaphysics and epistemology* (pp. 80–103). Oxford: Oxford University Press.
Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In Brownstein, M. and Saul, J. (Eds.), *Implicit bias and philosophy volume 1: Metaphysics and epistemology* (pp. 47–79). Oxford: Oxford University Press (Forthcoming in Brownstein and Saul, eds. Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology. Oxford: Oxford University Press)
Jennings, C., & Schwitzgebel, E. (2017). Women in philosophy: Quantitative analyses of specialization, prevalence, visibility, and generational change. *Public Affairs Quarterly*, *31*, 83–105.
Johnson, G. M. (2020). The structure of bias. *Mind*. https://doi.org/10.1093/mind/fzaa011.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The psychology of intuitive judgement* (1st ed., pp. 49–81). Cambridge: Cambridge University Press.
Kings, A. E. (2019). Philosophys diversity problem: Understanding the underrepresentation of women and minorities in philosophy. *Metaphilosophy*, *50*(3), 212–230.
Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801.
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.
Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (Clinical Research ed.)*, *296*(6623), 657.
Madva, A. (2016). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy*, *3*, 701.
Miconi, T. (2017). The impossibility of "fairness": a generalized impossibility result for decisions. arXiv:1707.01195.
Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology*, *85*(6), 1035–1048.
Munton, J. (2019a). Beyond accuracy: Epistemic flaws with statistical generalizations. *Philosophical Issues*, *29*(1), 228–240.
Munton, J. (2019b). Bias in a biased system: Visual perceptual prejudice. In *Bias, reason and enquiry: New perspectives from the crossroads of epistemology and psychology*. Oxford: Oxford University Press.
Nagel, M., Peppers-Bates, S., Leuschner, A., & Lindemann, A. (2018). Feminism and philosophy. *The American Philosophical Association*, *17*(2), 33.
Narayanan, A. (2016). Language necessarily contains human biases, and so will machines trained on language corpora. *Freedom to Tinker*. https://freedom-totinker.com/2016/08/24/language-necessarily-contains-human-biases-and-so-will-machines-trained-on-language-corpora/.
O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing Group.
Paxton, M., Figdor, C., & Tiberius, V. (2012). Quantifying the gender gap: An empirical study of the underrepresentation of women in philosophy. *Hypatia*, *27*(4), 949–957.
Price, R. (2016). *Microsoft is deleting its AI chatbot's incredibly racist tweets*. New York: Business Insider.
Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change?* (pp. 39–60). Oxford: Oxford University Press.
Soon, V. (2019). Implicit bias and social schema: A transactive memory approach. *Philosophical Studies*, 1–21.
Stephens-Davidowitz, S. (2014). *Opinion\Google, tell me. Is my son a genius?*. New York: The New York Times.
Stewart, A. J., & Valian, V. (2018). *An inclusive academy: Achieving diversity and excellence*. Cambridge: MIT Press.

Wilhelm, I., Conklin, S. L., & Hassoun, N. (2018). New data on the representation of women in philosophy journals: 20042015. *Philosophical Studies*, *175*(6), 1441–1464.
Wu, X. & Zhang, Z. (2016). Automated inference on criminality using face images. arXiv preprint arXiv:1611.04135.