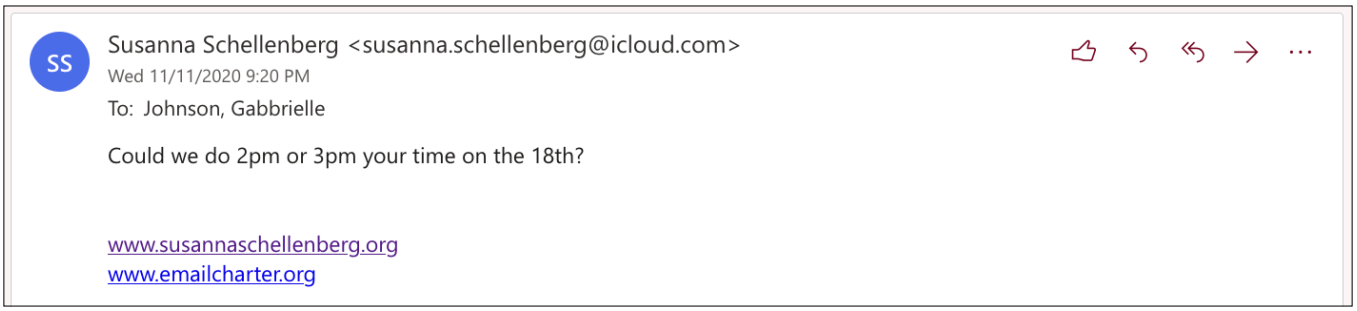
Analysis of NSF Grant Proposal Overlaps

Here I provide corroborating screenshots of my description of the events surrounding the creation of the Lebowitz joint document and Professor Schellenberg's use of my NSF grant proposal.

As evidenced by my email correspondences with Professor Schellenberg, she and I met three times over Zoom to discuss our Lebowitz Prize application.

First, on Nov. 18th, 2020, at 3pm PST as an initial consultation:



SS

Susanna Schellenberg <susanna.schellenberg@icloud.com>

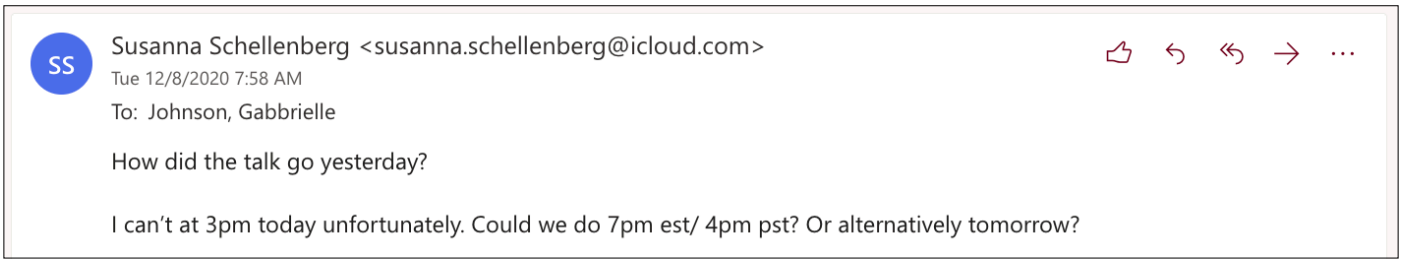
Wed 11/11/2020 9:20 PM

To: Johnson, Gabbrielle

Could we do 2pm or 3pm your time on the 18th?

www.susannaschellenberg.org
www.emailcharter.org

Second, on Dec. 8th, 2020, at 4pm PST to begin constructing the internal document that we would use only to circulate to our letter writers:



SS

Susanna Schellenberg <susanna.schellenberg@icloud.com>

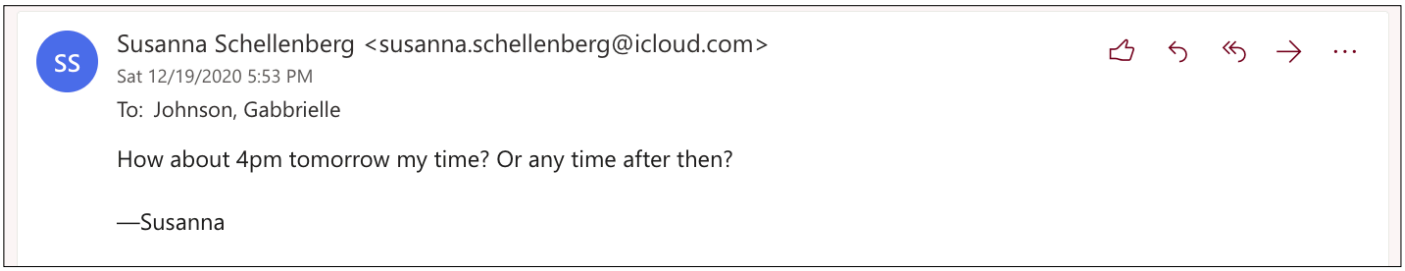
Tue 12/8/2020 7:58 AM

To: Johnson, Gabbrielle

How did the talk go yesterday?

I can't at 3pm today unfortunately. Could we do 7pm est/ 4pm pst? Or alternatively tomorrow?

And finally, on Dec. 20th, 2020, at 1pm PST, to finish that internal document:



SS

Susanna Schellenberg <susanna.schellenberg@icloud.com>

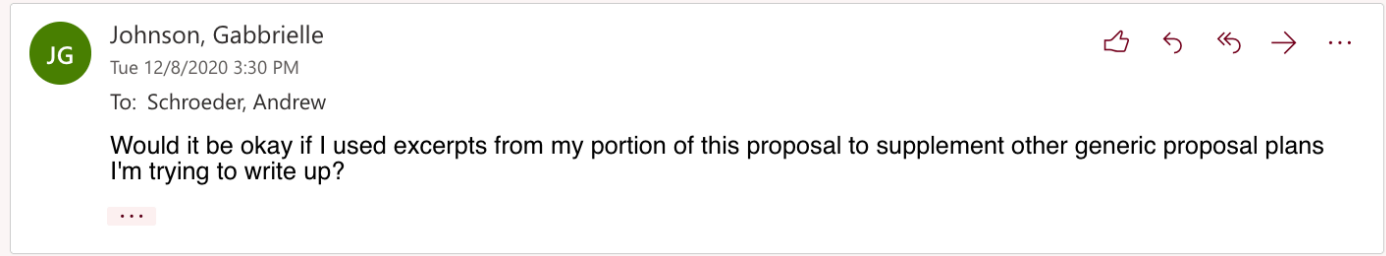
Sat 12/19/2020 5:53 PM

To: Johnson, Gabbrielle

How about 4pm tomorrow my time? Or any time after then?

—Susanna

Just before our meeting on Dec. 8th, 2020, at 4pm to create the internal document, I emailed my NSF grant proposal co-author Professor Andrew "Drew" Schroeder to ask for permission to use excerpts from my portion of our NSF grant draft for "generic proposal plans", having in mind this joint document:



By the time the meeting with Professor Schellenberg had started, I had not heard back from Professor Schroeder, but I had confidence he would give me permission. When the meeting started, I screen-shared my NSF grant proposal draft with Professor Schellenberg and told her I was thinking we could use excerpts from it for this internally circulated document, since we didn't plan to use it for the application. But I made clear that I was waiting on permission to use them from my co-author Professor Schroeder regardless. Professor Schellenberg told me that she also had pre-written materials that we could use solely for this document. So, I made a shared Google Doc online, where we each proceeded to copy and paste from our respective pre-written materials.

In what follows, I share screenshots of the edit history of that document to corroborate these points. For those not familiar, the dates along the right side indicate edits to the document in ascending order (with the first edits at the bottom and the subsequent edits rising to the top). Each edit is timestamped with the date and time the edit was made, as well as the user who made those edits. I am signed in under my account and so listed as "Gabbrielle Johnson" (with my edits in green); Professor Schellenberg is not signed in and so she is listed as "All anonymous users" (with her edits in gray).

In the first screenshot, you will see the first-ever edit to the document after its creation on December 8th, 2020, at 4:12pm. This edit, at 4:13pm, is me copying and pasting from the NSF grant proposal that I had just screen-shared with Professor Schellenberg, and that she had agreed we could copy and paste into this document to use solely for the purposes of creating an internally circulated document to our letter writers. As you can see from the copied and pasted material, all of the relevant verbatim overlap (and some additional verbatim overlap I just noticed) between my NSF grant proposal and Professor Schellenberg's *Phil Studies* submission exists in this initial copied and pasted material: (see page 3)

Fit

Only show named versions

First edit to the joint document at 4:13: me copying and pasting from the NSF grant draft

Decisions involving everyday citizens are increasingly guided by automated decision procedures, without any institutional mechanisms for ethical oversight. Recently, the tech industry has grown more concerned with the ethical ramifications of using these data-driven technologies, inviting input from philosophers, legal theorists, and other humanities scholars. However, such collaborations have largely been focused on the application of these technologies, such as their adaptation into existing legal frameworks and their effects on consumer privacy and autonomy. These scholars have only rarely been invited to discuss the internal workings of the automated decision procedures themselves – for example, to consider questions connected to labeling and quantifying data, formulating predictive models, and measuring bias and fairness. In our work, we aim to bridge the gap between traditional questions in philosophy (computational theories of mind) and contemporary issues concerning the application of machine learning programs.

There's a preconception that computer-based decision making is more objective and accurate than human decision making. But this conceals the moral and epistemic obstacles of computer-based decision making in at least two ways. Firstly, machine learning programs often instantiate so-called "black box" algorithms: those where it's arguably impossible for human programmers to describe the rationale for a particular outcome. Secondly, algorithmic biases rely on "proxy attributes": seemingly innocuous features that correlate with socially sensitive attributes, serving as proxies for the socially sensitive attributes themselves.

(Biases in Data)

Text overlap 18

An oft-cited motto in machine learning is, "Garbage in, garbage out". All machine learning programs are "trained" by identifying patterns in a known dataset and producing a predictive model dedicated to replicating those patterns for new data. If there are problematic patterns encoded in the training data, then its predictions will have those same problematic patterns.

Just-discovered textual overlap

Some issues arise from human bias in the data collection process. As has been highlighted in a number of news reports, automated facial recognition programs tend to identify women and people of color with lower accuracy than they identify white males. This stems from impoverished data collection practices: the images on which they are trained tend to be overpopulated with white male faces. As a result, the algorithms don't master recognition of features more prevalent in the faces of women and people of color. There are also biases that emerge from inexact data labeling. For example, every supervised learning program will be trained on data that is pre-labeled by humans. If the person labeling the data, perhaps due to personal prejudice, does so in a biased way, then we can expect inaccurate predictions. For example, an engineer might tend to rate male applicants for graduate school somewhat higher than equally-qualified female applicants. A machine learning program trained on this data will replicate that social prejudice.

Text overlap 17

What is less often discussed, however, is that similar problems can arise even in cases where data collection practices have no clear biases, because bias can emerge from ubiquitous and problematic social patterns. Imagine a graduate program in chemistry that uses an automated admissions program trained on past data about which students, when accepted, have gone on to successful careers in the field. If, as a matter of fact, men have historically been more successful in chemistry, the program will likely favor men in the application pool. But if men's disproportionate success in chemistry was a result of widespread personal or systemic injustices, then even careful data collection practices are apt to encode these societal injustices. Machine learning programs trained on such data will therefore perpetuate those unjust patterns.

Text overlap 11-15

An example of these problematic patterns can be seen in predictive policing algorithms like PredPol, which are used by law enforcement agencies across the country to identify potential "hotspots" for crime. Such algorithms rely on data in the form of historical records of the frequency of criminal activity in particular areas to make predictions about where police should be dispatched in anticipation of new crimes. Consider, then, what is apt to happen if historical policing practices have been shaped by discrimination. For example, if police have, due to historical patterns of racism, tended to over-patrol predominantly black and minority neighborhoods, then we can expect predictive software to continue to identify those neighborhoods as potential hotspots. It then dispatches police disproportionately to those areas, creating and collecting more data with which to continue the vicious cycle.

This is not simply an engineering problem. To avoid this sort of result, we must distinguish social patterns that reflect injustice from those that don't, and then we must figure out how to train machine learning programs either without using data reflecting unjust patterns, or else somehow compensating for the bias in such data. This will require a nuanced understanding of philosophical issues *and* extensive technical knowledge – inviting an interdisciplinary collaboration between philosophers and data scientists.

(Biases in Design)

When confronted with biased data, machine learning program engineers typically endeavor to create algorithms that operate on *just the facts*. However, this approach, even if applied successfully to accurate, representative data, cannot guarantee accurate predictions, since of course no algorithm is perfect. With every prediction, there is some risk of error. Data scientists must therefore decide how to manage this risk, and in particular must consider the importance of different types of error. Such decisions are implemented in part in the algorithmic design itself, e.g. in calibrating each algorithm's loss function.

For example, imagine you're tasked with building an image recognition program to distinguish human shapes from non-human shapes. The level of inaccuracy you should tolerate will depend on the use to which your algorithm will be put. If it is to be implemented in an office complex as a trigger to activate the automated lights, 75% accuracy would be inconvenient, but acceptable. However, if it is to be implemented in a self-driving car to prevent pedestrian collisions, you should demand near perfection. Algorithmic design decisions about how to manage error therefore inherently involve values.

The problem of how much error to tolerate gets much more complicated, however, because errors are not always distributed equally. A ProPublica exposé, for example, discovered the recidivism risk-assessment software COMPAS was over two-and-a-half times more likely to falsely flag black defendants as high-risk. Unfortunately, simple solutions – such as requiring algorithms to maintain comparable error rates in different groups – don't work, because equalizing one type of error may require distributing another type of error unequally. It is mathematically impossible, for example, to simultaneously equalize

December 8, 2020, 4:40 PM
All anonymous users

December 8, 2020, 4:39 PM
Gabbrielle Johnson

December 8, 2020, 4:38 PM
Gabbrielle Johnson

December 8, 2020, 4:37 PM
All anonymous users

December 8, 2020, 4:35 PM
All anonymous users

December 8, 2020, 4:34 PM
All anonymous users

December 8, 2020, 4:33 PM
All anonymous users

December 8, 2020, 4:32 PM
All anonymous users

December 8, 2020, 4:28 PM
All anonymous users

December 8, 2020, 4:26 PM
All anonymous users

December 8, 2020, 4:25 PM
All anonymous users

December 8, 2020, 4:24 PM
All anonymous users

December 8, 2020, 4:23 PM
All anonymous users

December 8, 2020, 4:22 PM
All anonymous users

December 8, 2020, 4:13 PM
Gabbrielle Johnson

December 8, 2020, 4:12 PM
Gabbrielle Johnson

Show changes

The second edit to the document (moving up the edit history on the right side of the screen), takes place at 4:22pm. This is Professor Schellenberg copying and pasting her pre-written materials into the document. As you can see from the copied and pasted material, most of this material overlaps with the material in our joint document as well as Professor Schellenberg's *Phil Studies* submission: (see page 5)

December 8, 2020, 4:22 ... Restore this version Version history

Fit

Only show named versions

Second edit at 4:22: Schellenberg copying and pasting from her materials.

Algorithmic Biases are Bottom-Up Biases not Top-Down Biases

We are living in the age of AI. Algorithms are used to make decisions about criminal sentencing, loans, policing, credit card applications, job recruiting, and medical care. What news we see on social media and which ads we are shown online are determined by algorithms. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the coronavirus crisis, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to negative consequences of AI.

All AI systems have biases. Many of these biases are deeply harmful and have harmful consequences for specific demographic groups. If one does on online search for a name predominantly given to black babies—Deshawn, Aisha, Jermaine—one is more likely to get an ad for criminal background checks than if one searches a name predominantly given to white babies—Geoffrey, Jill, and Emma.

Biases are typically discussed in the framework of top-down biases, that is, biases that stem from the effects that a person's beliefs, concepts, and background views have on her perceptions, thoughts, and actions. If someone has racist or sexist beliefs, this can affect how she interacts with people.

A computer programmer with such beliefs may design AI systems that exhibit the racist or sexist beliefs she harbors. Moreover, her background beliefs may affect how she interprets the data generated by an algorithm. There is ample evidence that such top-down biases exist in our perceptual and cognitive systems and that such top-down biases can affect how AI systems are designed and how their outputs are interpreted.

However, most biases in AI systems are not top-down biases. They stem not from the effects of a racist or sexist programmer's beliefs and concepts. They stem rather from how AI systems function at the lowest level: the patterns that the algorithm detects within the data it receives, the connections it forges between data points, and the generalizations that emerge from these connections. So they are not top-down biases, but rather bottom-up biases. While top-down biases are due to the goals, expectations, or beliefs of the programmer designing the algorithm, bottom-up biases are due to incoming data and its processing at the lowest level.

There are several reasons for why AI systems have bottom-up biases. One is that the incoming data can be biased. If an AI system operates on biased data, then it is hardly surprising that it will deliver biased results. After all, if the input is biased, then the output will be biased. Or as the old saying goes: "garbage in, garbage out".

A second reason is that AI systems operate on colossal and complex datasets and classify this data according to patterns it detects. They operate with statistical generalizations. The datasets consist of pieces of information, which we can call features. AI systems create vast feature spaces out of the incoming data. Algorithms pick up on regularities and correlations between features, link these features, and thereby create a feature space. The algorithm then projects that similar correlations between features will hold in the future and processes future data within the framework of this feature space. More specifically, it projects that the links held between features in the past will hold between features in the future.

In AI systems, these two sources of bottom-up biases can interact creating toxic feedback loops. A closer look at how online ads are generated will help see how. No programmer deviously wrote code so that ads for criminal background checks were prompted when a name typical in African American communities had been searched online. What happened is perhaps even more worrisome.

The initial source of the problem were the biases in the data on which the algorithm operated. The data was provided by us: the millions of people using google across the globe. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Enough of us must have done criminal background checks on certain kinds of names shortly after having searched for those names online, while at the same time rarely doing criminal background checks on other kind of names.

AdSense, the Google ad algorithm, detected this pattern, linked the relevant features, and then started to deliver ads suggestive of an arrest record after names common in African American communities had been searched online. The features linked may have included not only specific names. They may have also included proxy attributes, that is, attributes that correlate with some other feature in the environment such that one serves as a proxy for the other. In a racially segregated society, zip code is often a proxy for race. The ad algorithm picked up on the relevant correlations since it operates on statistical generalizations. This is a classic case of a bottom-up bias.

Once an AI system delivers ads in this manner it is no longer simply replicating the existing biases in our society. It is amplifying them. This paper analyzes the sources of bottom-up biases and investigates how these sources interact thereby amplifying biases in our society.

December 8, 2020, 4:40 PM
All anonymous users

December 8, 2020, 4:39 PM
Gabbrielle Johnson

December 8, 2020, 4:38 PM
Gabbrielle Johnson

December 8, 2020, 4:37 PM
All anonymous users

December 8, 2020, 4:35 PM
All anonymous users

December 8, 2020, 4:34 PM
All anonymous users

December 8, 2020, 4:33 PM
All anonymous users

December 8, 2020, 4:32 PM
All anonymous users

December 8, 2020, 4:28 PM
All anonymous users

December 8, 2020, 4:26 PM
All anonymous users

December 8, 2020, 4:25 PM
All anonymous users

December 8, 2020, 4:24 PM
All anonymous users

December 8, 2020, 4:23 PM
All anonymous users

December 8, 2020, 4:22 PM
All anonymous users

December 8, 2020, 4:13 PM
Gabbrielle Johnson

December 8, 2020, 4:12 PM
Gabbrielle Johnson

Show changes

We then begin editing this document together. However, given that I still had not received permission from Professor Schroeder to use the excerpts from the NSF grant proposal, I suggested to Professor Schellenberg that we leave my proposal excerpts untouched at the bottom of the shared document, waiting to incorporate them until I got permission from Professor Schroeder. I did not receive permission before our initial meeting ended, and so we never in that meeting incorporated parts of the NSF grant proposal into the internal document. We continued to work on the document together until Professor Schellenberg had something come up that required she leave the meeting.

On December 9th, 2020, I hear back from Professor Schroeder, where he tells me that I can use the materials, but to let him know if at any point I intend to use them for an official application. I agree on December 11th, 2020, to not use any substantial excerpts for any official applications, again knowing that Professor Schellenberg and I had agreed to use the excerpts only for the purposes of circulating the document internally to our letter writers:

SA

Schroeder, Andrew

Wed 12/9/2020 10:40 AM

To: Johnson, Gabbrielle

Definitely okay to use shorter extracts (a few sentences to a paragraph) or not-clearly-recognizable longer extracts (using the same ideas/structure, but paraphrasing or reordering). Probably okay to use longer extracts, too – but please check with me before sending off in any official application. I would just want to make sure no funding agency was going to have your thing and Hiram's thing on their desk at the same time.

FYI, Hiram did send off the draft to the NSF guy for his impressions. Will let you know when we hear back.

...

JG

Johnson, Gabbrielle

Fri 12/11/2020 12:20 PM

To: Schroeder, Andrew

Sorry, I thought I had responded to this: thanks, that aligns with what I was thinking. I'll def let you know if I feel I need to use substantial excerpts for any official applications, but I don't think that will be necessary. I can rewrite most portions for those purposes. I just want to use the example of the light switch for some other stuff and some very nearby language of the relationship between philosophy and engineers. Nothing excessive.

And good to know that Hiram thought it was in a position to forward. Excited to hear what the NSF guy thinks.

Question: are we discussing the PPA revisions this afternoon or has that been kicked down the line again?

...

Professor Schellenberg and I meet again over Zoom on December, 20th, 2020, at 1pm, beginning work again on the shared Google Doc we had worked on last time: (see page 7)

Fit

We begin working on the document again after I've received permission from Schroeder (now 12/20/20)

Are Algorithmic Biases Bottom-Up or Top-Down?

We are living in the age of AI. Algorithms are used to make decisions about criminal sentencing, loans, policing, credit card applications, job recruiting, and medical care. What news we see on social media and which ads we are shown online are determined by algorithms. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the coronavirus crisis, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to negative consequences of AI.

All AI systems have biases. Many of these biases are deeply harmful and have harmful consequences for specific demographic groups. If one does an online search for a name predominantly given to black babies—Deshawn, Aisha, Jermaine—one is more likely to get an ad for criminal background checks than if one searches a name predominantly given to white babies—Geoffrey, Jill, and Emma.

Biases are typically discussed in the framework of top-down biases, that is, biases that stem from the effects that a person's beliefs, concepts, and background views have on her perceptions, thoughts, and actions. If someone has racist or sexist beliefs, this can affect how she interacts with people.

A computer programmer with such beliefs may design AI systems that exhibit the racist or sexist beliefs she harbors. Moreover, her background beliefs may affect how she interprets the data generated by an algorithm. There is ample evidence that such top-down biases exist in our perceptual and cognitive systems and that such top-down biases can affect how AI systems are designed and how their outputs are interpreted.

However, we both agree that most biases in AI systems are not top-down biases. They stem not from the effects of a racist or sexist programmer's beliefs and concepts. They stem rather from how AI systems function at the lowest level: the patterns that the algorithm detects within the data it receives, the connections it forges between data points, and the generalizations that emerge from these connections. So they are not top-down biases, but rather bottom-up biases. While top-down biases are due to the goals, expectations, or beliefs of the programmer designing the algorithm, bottom-up biases are due to incoming data and its processing at the lowest level.

While we agree that most biases in AI systems are bottom-up, we disagree as to what the relation if any there is between bottom-up and top-down biases. Johnson argues that bottom-up and top-down biases are intrinsically intertwined. By contrast, Schellenberg argues that at least some bottom-up biases are independent of any top-down biases. So while Schellenberg has it that at least some biases are pure bottom-up biases, Johnson holds that pure bottom-up biases do not exist. As she argues all bottom-up biases are better

[interpreted as low-level instantiations of individual-level biases; they are what they are in virtue of top-down causal interactions].

Statistical regularities are not top-down biases.

Johnson's view: There is no such thing as a pure bottom-up bias because what grounds explanations of why features are linked via low-level processes (i.e., why bottom-up biases exist) ultimately depends on individual-level descriptions of a subject's interaction with the wider environment. For these reasons, one can't identify bottom-up biases without reference to the causal-explanatory connections between the whole subject and the wider environmental regularities with which they are attuned. Slogan: the natures of bottom-up biases depend constitutively on top-down influences.

Why does this difference matter? --- SEE BELOW p. 4

/// We can all agree that top-down biases are morally problematic.

Schellenberg holds that pure bottom-up biases can be morally problematic as well. While there is ///

Often bottom-biases seem utterly innocuous -- even while they have pernicious and racist and sexist effects.

[[The differences between these two approaches can be explained more clearly by taking a closer look at different types of bottom-up biases.

There are several reasons for why AI systems have bottom-up biases. One is that the incoming data can be biased. If an AI system operates on biased data, then it is hardly surprising that it will deliver biased results. After all, if the input is biased, then the output will be biased. Or as the old saying goes: "garbage in, garbage out".

A second reason is that AI systems operate on colossal and complex datasets and classify this data according to patterns it detects. They operate with statistical generalizations. The datasets consist of pieces of information, which we can call features. AI systems create vast feature spaces out of the incoming data. Algorithms pick up on regularities and correlations between features, link these features, and thereby create a feature space. The algorithm then projects that similar correlations between features will hold in the future and processes future data within the framework of this feature space. More specifically, it projects that the links held between features in the past will hold between features in the future.]]

Only show named versions

December 20, 2020, 1:48 PM
All anonymous users

December 20, 2020, 1:46 PM
Gabbrielle Johnson

December 20, 2020, 1:46 PM
All anonymous users

December 20, 2020, 1:45 PM
All anonymous users

December 20, 2020, 1:45 PM
All anonymous users

December 20, 2020, 1:44 PM
All anonymous users

December 20, 2020, 1:44 PM
All anonymous users

December 20, 2020, 1:42 PM
All anonymous users

December 20, 2020, 1:40 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:38 PM
All anonymous users

December 20, 2020, 1:38 PM
All anonymous users

December 20, 2020, 1:37 PM
All anonymous users

December 20, 2020, 1:37 PM
All anonymous users

Show changes

I tell her I got permission from my coauthor to use excerpts from the NSF grant proposal so long as we only use it for internally circulating to our letter writers. It is at this point that Professor Schellenberg, at 1:40pm, copies my portions of the NSF grant proposal that are still below the main text, and pastes them to the top of the document for the purposes of now integrating those excerpts into the main exposition: (see page 9)

Fit

Only show named versions

Schellenberg copying and pasting excerpts from my grant at the top of the document.

Text overlap 18

Just-discovered textual overlap

Text overlap 17

Text overlap 11-15

Are Algorithmic Biases Bottom-Up or Top-Down?

Begin Spiel:
An oft-cited motto in machine learning is, "Garbage in, garbage out". All machine learning programs are "trained" by identifying patterns in a known dataset and producing a predictive model dedicated to replicating those patterns for new data. If there are problematic patterns encoded in the training data, then its predictions will have those same problematic patterns.

Some issues arise from human bias in the data collection process. As has been highlighted in a number of news reports, automated facial recognition programs tend to identify women and people of color with lower accuracy than they identify white males. This stems from impoverished data collection practices: the images on which they are trained tend to be overpopulated with white male faces. As a result, the algorithms don't master recognition of features more prevalent in the faces of women and people of color. There are also biases that emerge from inexact data labeling. For example, every supervised learning program will be trained on data that is pre-labeled by humans. If the person labeling the data, perhaps due to personal prejudice, does so in a biased way, then we can expect inaccurate predictions. For example, an engineer might tend to rate male applicants for graduate school somewhat higher than equally-qualified female applicants. A machine learning program trained on this data will replicate that social prejudice.

What is less often discussed, however, is that similar problems can arise even in cases where data collection practices have no clear biases, because bias can emerge from ubiquitous and problematic social patterns. Imagine a graduate program in chemistry that uses an automated admissions program trained on past data about which students, when accepted, have gone on to successful careers in the field. If, as a matter of fact, men have historically been more successful in chemistry, the program will likely favor men in the application pool. But if men's disproportionate success in chemistry was a result of widespread personal or systemic injustices, then even careful data collection practices are apt to encode these societal injustices. Machine learning programs trained on such data will therefore perpetuate those unjust patterns.

An example of these problematic patterns can be seen in predictive policing algorithms like PredPol, which are used by law enforcement agencies across the country to identify potential "hotspots" for crime. Such algorithms rely on data in the form of historical records of the frequency of criminal activity in particular areas to make predictions about where police should be dispatched in anticipation of new crimes. Consider, then, what is apt to happen if historical policing practices have been shaped by discrimination. For example, if police have, due to historical patterns of racism, tended to over-patrol predominantly black and minority neighborhoods, then we can expect predictive software to continue to identify those neighborhoods as potential hotspots. It then dispatches police disproportionately to those areas, creating and collecting more data with which to continue the vicious cycle.

This is not simply an engineering problem. To avoid this sort of result, we must distinguish computational patterns that reflect injustice from those that don't, and then we must figure out how to train machine learning programs either without using data reflecting unjust patterns, or else somehow compensating for the bias in such data. This will require a nuanced understanding of philosophical issues *and* extensive technical knowledge – inviting an interdisciplinary collaboration between philosophers and data scientists.

We are living in the age of AI. Algorithms are used to make decisions about criminal sentencing, loans, policing, credit card applications, job recruiting, and medical care. What news we see on social media and which ads we are shown online are determined by algorithms. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the coronavirus crisis, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to negative consequences of AI.

All AI systems have biases. Many of these biases are deeply harmful and have harmful consequences for specific demographic groups. If one does an online search for a name predominantly given to black babies—Deshawn, Aisha, Jermaine—one is more likely to get an ad for criminal background checks than if one searches a name predominantly given to white babies—Geoffrey, Jill, and Emma.

Biases are typically discussed in the framework of top-down biases, that is, biases that stem from the effects that a person's beliefs, concepts, and background views have on her perceptions, thoughts, and actions. If someone has racist or sexist beliefs, this can affect how she interacts with people.

A computer programmer with such beliefs may design AI systems that exhibit the racist or sexist beliefs she harbors. Moreover, her background beliefs may affect how she interprets the data generated by an algorithm. There is ample evidence that such top-down biases exist in our perceptual and cognitive systems and that such top-down biases can affect how AI systems are designed and how their outputs are interpreted.

However, we both agree that most biases in AI systems are not top-down biases. They stem not from the effects of a racist or sexist programmer's beliefs and concepts. They stem rather from how AI systems function at the lowest level: the patterns that the algorithm detects within the data it receives, the connections it forges between data points, and the generalizations that emerge from these connections. So they are not top-down biases, but rather bottom-up biases. While top-down biases are due to the goals, expectations, or beliefs of the programmer designing the algorithm, bottom-up biases are due to incoming data and its processing at the lowest level.

While we agree that most biases in AI systems are bottom-up, we disagree as to what the relation if any there

December 20, 2020, 1:48 PM
All anonymous users

December 20, 2020, 1:46 PM
Gabbrielle Johnson

December 20, 2020, 1:46 PM
All anonymous users

December 20, 2020, 1:45 PM
All anonymous users

December 20, 2020, 1:45 PM
All anonymous users

December 20, 2020, 1:44 PM
All anonymous users

December 20, 2020, 1:44 PM
All anonymous users

December 20, 2020, 1:42 PM
All anonymous users

December 20, 2020, 1:40 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:38 PM
All anonymous users

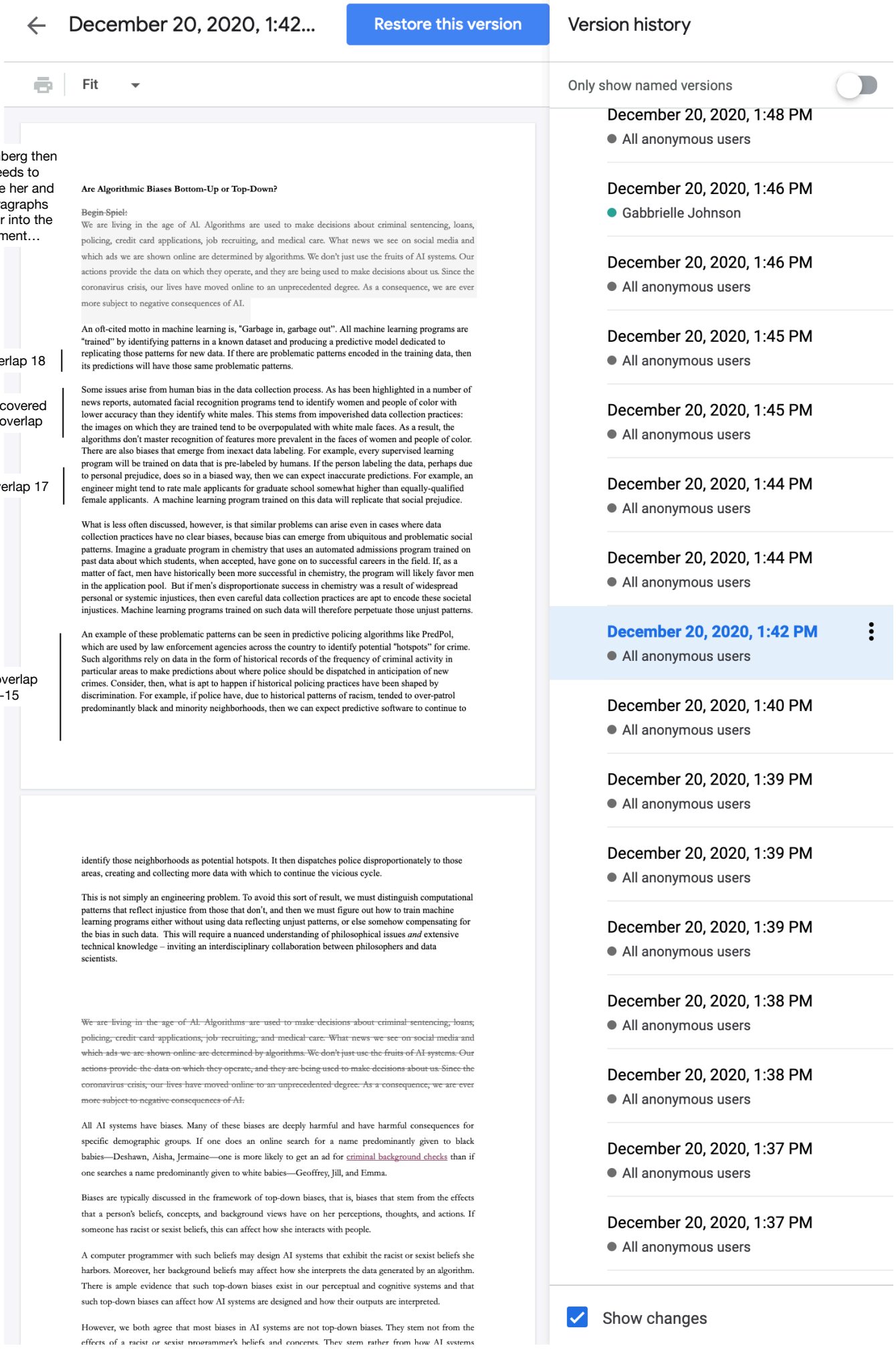
December 20, 2020, 1:38 PM
All anonymous users

December 20, 2020, 1:37 PM
All anonymous users

December 20, 2020, 1:37 PM
All anonymous users

Show changes

She then proceeds to, over the next several edits, integrate my excerpts in between paragraphs of the main text of the document. Again, you can see from the screen shots to follow that these include the intact verbatim overlap passages that end up in her *Phil Studies* submission: (see pages 11 and 12)


December 20, 2020, 1:42...
Restore this version
Version history
Fit
Only show named versions
Schellenberg then proceeds to integrate her and my paragraphs together into the document…
Are Algorithmic Biases Bottom-Up or Top-Down?
~~Begin Spiel:~~
We are living in the age of AI. Algorithms are used to make decisions about criminal sentencing, loans, policing, credit card applications, job recruiting, and medical care. What news we see on social media and which ads we are shown online are determined by algorithms. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the coronavirus crisis, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to negative consequences of AI.
An oft-cited motto in machine learning is, "Garbage in, garbage out". All machine learning programs are "trained" by identifying patterns in a known dataset and producing a predictive model dedicated to replicating those patterns for new data. If there are problematic patterns encoded in the training data, then its predictions will have those same problematic patterns.
Text overlap 18
Just-discovered textual overlap
Some issues arise from human bias in the data collection process. As has been highlighted in a number of news reports, automated facial recognition programs tend to identify women and people of color with lower accuracy than they identify white males. This stems from impoverished data collection practices: the images on which they are trained tend to be overpopulated with white male faces. As a result, the algorithms don't master recognition of features more prevalent in the faces of women and people of color. There are also biases that emerge from inexact data labeling. For example, every supervised learning program will be trained on data that is pre-labeled by humans. If the person labeling the data, perhaps due to personal prejudice, does so in a biased way, then we can expect inaccurate predictions. For example, an engineer might tend to rate male applicants for graduate school somewhat higher than equally-qualified female applicants. A machine learning program trained on this data will replicate that social prejudice.
Text overlap 17
What is less often discussed, however, is that similar problems can arise even in cases where data collection practices have no clear biases, because bias can emerge from ubiquitous and problematic social patterns. Imagine a graduate program in chemistry that uses an automated admissions program trained on past data about which students, when accepted, have gone on to successful careers in the field. If, as a matter of fact, men have historically been more successful in chemistry, the program will likely favor men in the application pool. But if men's disproportionate success in chemistry was a result of widespread personal or systemic injustices, then even careful data collection practices are apt to encode these societal injustices. Machine learning programs trained on such data will therefore perpetuate those unjust patterns.
Text overlap 11-15
An example of these problematic patterns can be seen in predictive policing algorithms like PredPol, which are used by law enforcement agencies across the country to identify potential "hotspots" for crime. Such algorithms rely on data in the form of historical records of the frequency of criminal activity in particular areas to make predictions about where police should be dispatched in anticipation of new crimes. Consider, then, what is apt to happen if historical policing practices have been shaped by discrimination. For example, if police have, due to historical patterns of racism, tended to over-patrol predominantly black and minority neighborhoods, then we can expect predictive software to continue to
identify those neighborhoods as potential hotspots. It then dispatches police disproportionately to those areas, creating and collecting more data with which to continue the vicious cycle.
This is not simply an engineering problem. To avoid this sort of result, we must distinguish computational patterns that reflect injustice from those that don't, and then we must figure out how to train machine learning programs either without using data reflecting unjust patterns, or else somehow compensating for the bias in such data. This will require a nuanced understanding of philosophical issues *and* extensive technical knowledge – inviting an interdisciplinary collaboration between philosophers and data scientists.
~~We are living in the age of AI. Algorithms are used to make decisions about criminal sentencing, loans, policing, credit card applications, job recruiting, and medical care. What news we see on social media and which ads we are shown online are determined by algorithms. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the coronavirus crisis, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to negative consequences of AI.~~
All AI systems have biases. Many of these biases are deeply harmful and have harmful consequences for specific demographic groups. If one does an online search for a name predominantly given to black babies—Deshawn, Aisha, Jermaine—one is more likely to get an ad for criminal background checks than if one searches a name predominantly given to white babies—Geoffrey, Jill, and Emma.
Biases are typically discussed in the framework of top-down biases, that is, biases that stem from the effects that a person's beliefs, concepts, and background views have on her perceptions, thoughts, and actions. If someone has racist or sexist beliefs, this can affect how she interacts with people.
A computer programmer with such beliefs may design AI systems that exhibit the racist or sexist beliefs she harbors. Moreover, her background beliefs may affect how she interprets the data generated by an algorithm. There is ample evidence that such top-down biases exist in our perceptual and cognitive systems and that such top-down biases can affect how AI systems are designed and how their outputs are interpreted.
However, we both agree that most biases in AI systems are not top-down biases. They stem not from the effects of a racist or sexist programmer's beliefs and concepts. They stem rather from how AI systems
December 20, 2020, 1:48 PM
All anonymous users
December 20, 2020, 1:46 PM
Gabbrielle Johnson
December 20, 2020, 1:46 PM
All anonymous users
December 20, 2020, 1:45 PM
All anonymous users
December 20, 2020, 1:45 PM
All anonymous users
December 20, 2020, 1:44 PM
All anonymous users
December 20, 2020, 1:44 PM
All anonymous users
December 20, 2020, 1:42 PM
All anonymous users
December 20, 2020, 1:40 PM
All anonymous users
December 20, 2020, 1:39 PM
All anonymous users
December 20, 2020, 1:39 PM
All anonymous users
December 20, 2020, 1:39 PM
All anonymous users
December 20, 2020, 1:38 PM
All anonymous users
December 20, 2020, 1:38 PM
All anonymous users
December 20, 2020, 1:37 PM
All anonymous users
December 20, 2020, 1:37 PM
All anonymous users
Show changes

Fit

Only show named versions

…folding paragraphs from each into the main document.

Are Algorithmic Biases Bottom-Up or Top-Down?

We are living in the age of AI. Algorithms are used to make decisions about criminal sentencing, loans, policing, credit card applications, job recruiting, and medical care. What news we see on social media and which ads we are shown online are determined by algorithms. We don't just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the coronavirus crisis, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to negative consequences of AI.

Text overlap 18

An oft-cited motto in AI is, "Garbage in, garbage out". All AI programs are "trained" by identifying patterns in a known dataset and producing a predictive model dedicated to replicating those patterns for new data. If there are problematic patterns encoded in the training data, then its predictions will have those same problematic patterns.

Just-discovered textual overlap

Some issues arise from human bias in the data collection process. As has been highlighted in a number of news reports, automated facial recognition programs tend to identify women and people of color with lower accuracy than they identify white males. This stems from impoverished data collection practices: the images on which they are trained tend to be overpopulated with white male faces. As a result, the algorithms don't master recognition of features more prevalent in the faces of women and people of color. There are also biases that emerge from inexact data labeling. For example, every supervised learning program will be trained on data that is pre-labeled by humans. If the person labeling the data, perhaps due to personal prejudice, does so in a biased way, then we can expect inaccurate predictions. For example, an engineer might tend to rate male applicants for graduate school somewhat higher than equally-qualified female applicants. A machine learning program trained on this data will replicate that social prejudice.

Text overlap 17

Biases are typically discussed in the framework of top-down biases, that is, biases that stem from the effects that a person's beliefs, concepts, and background views have on her perceptions, thoughts, and actions. If someone has racist or sexist beliefs, this can affect how she interacts with people.

A computer programmer with such beliefs may design AI systems that exhibit the racist or sexist beliefs she harbors. Moreover, her background beliefs may affect how she interprets the data generated by an algorithm. There is ample evidence that such top-down biases exist in our perceptual and cognitive systems and that such top-down biases can affect how AI systems are designed and how their outputs are interpreted.

However, we both agree that most biases in AI systems are not top-down biases. They stem not from the effects of a racist or sexist programmer's beliefs and concepts. They stem rather from how AI systems function at the lowest level: the patterns that the algorithm detects within the data it receives, the connections it forges between data points, and the generalizations that emerge from these connections. So they are not top-down biases, but rather bottom-up biases. While top-down biases are due to the goals, expectations, or beliefs of the programmer designing the algorithm, bottom-up biases are due to incoming data and its processing at the lowest level.

While we agree that most biases in AI systems are bottom-up, we disagree as to what the relation if any there is between bottom-up and top-down biases. Johnson argues that bottom-up and top-down biases are intrinsically intertwined. By contrast, Schellenberg argues that at least some bottom-up biases are independent of any top-down biases. So while Schellenberg has it that at least some biases are pure bottom-up biases, Johnson holds that pure bottom-up biases do not exist. As she argues all bottom-up biases are better [interpreted as low-level instantiations of individual-level biases; they are what they are in virtue of top-down causal interactions].

What is less often discussed, however, is that similar problems can arise even in cases where data collection practices have no clear biases, because bias can emerge from ubiquitous and problematic social patterns. Imagine a graduate program in chemistry that uses an automated admissions program trained on past data about which students, when accepted, have gone on to successful careers in the field. If, as a matter of fact, men have historically been more successful in chemistry, the program will likely favor men in the application pool. But if men's disproportionate success in chemistry was a result of widespread personal or systemic injustices, then even careful data collection practices are apt to encode these societal injustices. Machine learning programs trained on such data will therefore perpetuate those unjust patterns.

Text overlap 11-15

An example of these problematic patterns can be seen in predictive policing algorithms like PredPol, which are used by law enforcement agencies across the country to identify potential "hotspots" for crime. Such algorithms rely on data in the form of historical records of the frequency of criminal activity in particular areas to make predictions about where police should be dispatched in anticipation of new crimes. Consider, then, what is apt to happen if historical policing practices have been shaped by discrimination. For example, if police have, due to historical patterns of racism, tended to over-patrol predominantly black and minority neighborhoods, then we can expect predictive software to continue to identify those neighborhoods as potential hotspots. It then dispatches police disproportionately to those areas, creating and collecting more data with which to continue the vicious cycle.

This is not simply an engineering problem. To avoid this sort of result, we must distinguish computational patterns that reflect injustice from those that don't, and then we must figure out how to train machine learning programs either without using data reflecting unjust patterns, or else somehow compensating for the bias in such data. This will require a nuanced understanding of philosophical issues *and* extensive technical knowledge – inviting an interdisciplinary collaboration between philosophers and data scientists.

December 20, 2020, 1:48 PM
All anonymous users

December 20, 2020, 1:46 PM
Gabbrielle Johnson

December 20, 2020, 1:46 PM
All anonymous users

December 20, 2020, 1:45 PM
All anonymous users

December 20, 2020, 1:45 PM
All anonymous users

December 20, 2020, 1:44 PM
All anonymous users

December 20, 2020, 1:44 PM
All anonymous users

December 20, 2020, 1:42 PM
All anonymous users

December 20, 2020, 1:40 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:39 PM
All anonymous users

December 20, 2020, 1:38 PM
All anonymous users

December 20, 2020, 1:38 PM
All anonymous users

December 20, 2020, 1:37 PM
All anonymous users

December 20, 2020, 1:37 PM
All anonymous users

Show changes

On January 5th, 2021, our Lebowitz Prize application is submitted. The submitted materials include two short abstracts. Fitting with my agreement with Professor Schroeder and my agreement with Professor Schellenberg, none of the submitted materials contain any of the language from the NSF grant proposal.