Relevant Background: Johnson claims that I took from the rough draft of her application for an NSF grant, despite the fact that she has never shared any document relating to her NSF application with me or even mentioned to me that she is applying for a NSF grant. I saw the draft that she sent to *Philosophical Studies* for the first time on September 27, 2021 when it was sent to me by the Editors-in-Chief. The draft sent to me is not available online. I could not possibly have had access to it. I do not know on what grounds Johnson claims that I could possibly have taken from it.

The relevant passages in my submission that overlap with Johnson's NSF application are in the 5-page unpublished document that Johnson and I co-authored below. For ease of assessment the relevant passages are highlighted (see pages 1 and 2) and numbered in accordance with how the instances of textual overlap are numbered in the letter from the Editors-in-Chief of Philosophical Studies sent to Rutgers University on November 5, 2021 (textual overlap 11-15, 17, 18).

Johnson does not mention this 5-page document in either her September 27, 2021 communication, nor in her October 11, 2021 letter to the Editors-in-Chief of *Philosophical Studies*. In her October 11, 2021 letter she does not respond to the fact that—as laid out in my October 1, 2021 letter—I have never seen this (or any other) draft of her application for an NSF grant. She does not mention her rough draft of an application for a NSF grant at all in her October 11 letter. I do not know why Johnson claims I took from a draft of an application for a NSF grant that she knows I have never seen.

Explanation for why this 5-page document was written: In December 2020, I suggested to Johnson that we apply for the Lebowitz Prize. This prize is given "to a pair of philosophers who hold contrasting (not necessarily opposing) views of an important philosophical question that is of current interest both to the field and to an educated public audience. … In selecting the winners, the committee is encouraged to consider candidates who address important philosophical issues using diverse methodologies and perspectives and who, in so doing, are able to communicate important philosophical ideas to specialist and non-specialist audiences."[1] The two philosophers must be (self-)nominated jointly. The application for the prize includes only the CV of each nominee and a 300-word abstract each that "should be tailored to address the specific topic to be discussed as well as the approach of the speaker".

Since our letter writers were not familiar with the literature on algorithmic bias and since a 300-word abstract was very thin material on the basis of which to write a letter, Johnson and I wrote the following 5-page document so as to make it easier for them to write their letters. This was the sole purpose of this unpublished document. We did not win the prize and agreed that the material of our joint application can be used freely by each of us.

I am the primary author of pages 1-3 of this unpublished document (most of it comes straight from a new paper on biased algorithms that I was writing at the time and I have since discarded while integrating the material in my submission to *Philosophical Studies*). Johnson is the sole author of pages 4-5. Since I am the primary author of pages 1-3 and since we agreed that the material in our joint application can be used freely by each of us, I used the material in pages 1-3 material in my submission. However, on careful inspection, I see that Johnson did provide some material that made it into pages 1-3. This is the material referred to as textual overlap 11- 15 and 17 from the November 5, 2021 letter sent to Rutgers University by the Editors-in-Chief of *Philosophical Studies*. I apologize profusely if there was poor communication about us using the material in our joint application freely by each of us and for material that Johnson contributed to this 5-page unpublished document ending up in my submission. This was an honest mistake. To my knowledge, textual overlap between a single authored unpublished submission and a jointly authored unpublished document does not qualify as plagiarism on any understanding of the term (assuming the author of the single authored document is one of the authors of the jointly authored document).

The applications for this prize are assessed by a committee that may not involve any experts. For this reason, the document is written such that the issues are understandable to a wide audience unfamiliar with the debate. Thus, I present the issues in the simple terms I use in my work and leave out the big differences between Johnson and my views.

[1] For details, see https://www.apaonline.org/page/lebowitz

Algorithmic Bias

Gabbrielle Johnson and Susanna Schellenberg

We are living in the age of Al. 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 do not only use the fruits of AI systems. Our actions provide the data on which they operate. Based on this data, algorithms are being used to make decisions about us that range from the superficial to the life changing. Since the pandemic, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to the positive and negative consequences of AI.

Like natural intelligent systems, all AI systems have biases. The reason is that there is a mismatch in data and processing power. To deal with this mismatch, such systems use statistical generalization to operate efficiently. The human visual system, for instance, is riddled with biases. Some of these biases, such as the bias that light comes from above, are beneficial. Some are harmful. Given how flawed humans are, it may not be surprising that we are biased. After all, we are driven by fears, act on the basis of emotions, and are prone to misunderstandings.

It may, however, come as a surprise that AI systems are biased. After all, computer algorithms form the core of AI systems and these algorithms are grounded in mathematics and operate on raw data. So one might expect them to be objective and just. But they are not. 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. As the oft-cited motto goes, "garbage in, garbage out".

Text overlap 18

Some problematic patterns come to be encoded in the training data due to 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 do not 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 a white applicant for graduate school somewhat higher than equally-qualified black applicant. An AI 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, as we argue, most biases in AI systems are not top-down biases. Most biases 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.

Text overlap

11 An example of a bottom-up bias is the predictive policing algorithm PredPol. It is used by law
12 enforcement agencies across the country to identify potential "hotspots" for crime. Its algorithm operates on data that consists of historical records of the frequency of criminal activity in particular
13 areas. On the basis of this historical data, the algorithm makes predictions about where police should
14 be dispatched in anticipation of new crimes. Consider, then, what is apt to happen if historical policing
15 practices have been shaped by discrimination. If police have, due to historical patterns of racism, tended to over-patrol predominantly black 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.

***

While we agree that most biases in AI systems are bottom-up, we disagree as to what 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].

These two ways of understanding bottom-up biases have ramifications on how algorithms are evaluated and what interventions are likely to be successful in making the algorithm have less harmful effects. The two understandings make different prescriptions for when and how to intervene.

According to Schellenberg's view of bottom-up biases, it is possible to make an algorithm less biased simply by improving on the algorithm itself: intervening on the computational processes that give rise to them: the manipulation of training data, input filters, and blocking transformation processes, to name just a few examples.

By contrast, according to Johnson's view, this approach will never be successful: if all bottom-up biases are intertwined with top-down biases, any ameliorative interventions will require intervening on wider societal patterns that give rise to the bottom-up biases.

Both approaches have their virtues and vices: computational interventions are easier to achieve, but short-lived. Societal interventions are hard to bring about, but once successful long-lasting. While there are deep differences between our two ways of understanding bottom-up biases, there is significant common ground, especially when it comes to strategies for making algorithms less biased.

Johnson acknowledges that interventions that affect only the algorithm, absent interventions on the wider patterns of societal injustice, are our best practical approach to curbing algorithmic bias. After all, such interventions pinpoint a precise location in a complex causal network at which we can intervene. Schellenberg acknowledges that many bottom-up biases are not pure and that changing societal patterns that underpin such bottom-up biases will make it possible to eliminate that bias. Moreover, she acknowledges that even pure bottom-up biases interact with societal patterns and that changing such societal patterns will affect and perhaps even eliminate the algorithmic bias. We both believe that given the ever-growing feedback loops that exist between humans and machines, interventions on algorithms is necessary to dismantle biases harbored by humans and machines alike.

More Technical Summary:

- Standard Bias Case: Fx → Gx, where F is a social group and G is a property stereotypical of that social group. (Imagine: Black people are dangerous.)

Question: Why does the system transition from F to G?

Two contrasting approaches:

1. Bottom-up approach (Schellenberg): Features F-G are linked, because they're statistically correlated in the data.
2. Top-down approach (Johnson): There's a social stereotype that Fs are G, because of causal-discriminatory patterns in the environment.

Divergences:

I. Moral Evaluation

We both agree that Standard Bias Case is morally problematic from both directions. But more interesting cases emerge where moral evaluation might diverge depending on the approach:

- Proxy Bias Case: Hx → Gx, where H is NOT a social group, but correlates with one, and G is a property stereotypical of that social group. (Imagine: people in zip code 90011 are dangerous.)

1. According to the Bottom-up approach, this is not necessarily problematic because the linked features H-G do not include reference to social kind membership (indeed, the data might not include social kinds at all).
2. According to the top-down approach, this is potentially problematic because it's ultimately still the stereotype that Fs are G [that explains why the system transitions from H to G], because F and H correlate in the wider environment.

II. Epistemic Evaluation

- Proxy Bias Case: Hx → Gx, where H is NOT a social group, but correlates with one, and G is a property stereotypical of that social group. (Imagine: people with criminal records are dangerous.

1. According to the bottom-up approach, this is epistemically kosher so long as H really is (statistically) good evidence of G.
2. According to the top-down approach, epistemically problematic because it's ultimately still the stereotype that Fs are G [that explains why the system transitions from H to G].

III. Intervention

1. According to the bottom-up approach, there are three ways to break down a problematic bias between F and G: to disrupt statistical patterns in the data, eliminate access to feature F in the inputs, or introduce a filter that prohibits transitioning to G from F.
2. According to the top-down approach, in all of these cases, the bias Fs are G will emerge, because the primary cause is the social stereotype that Fs are G in the wider social environment. Only when we disrupt this social stereotype will we truly eliminate the bias Fs are G from the system.