RUTGERS
School of Arts and Sciences

Susanna Schellenberg
Distinguished Professor of Philosophy
Department of Philosophy
Rutgers, The State University of New Jersey
106 Somerset Street, 5th Floor
New Brunswick, NJ 08901-4800

Web: http://philosophy.rutgers.edu
Phone: 732-932-9861
Fax: 732-932-8617

November 16, 2021

Dear Editors,

On November 5, 2021, you sent notice of allegations of plagiarism related to my unpublished paper "Varieties of Biased Algorithms" that I was invited to submit for inclusion in a special issue of *Philosophical Studies* (hereafter, "the submission") to several parties at Rutgers University (hereafter "the Editors' letter"). I am deeply distressed by these allegations. I have previously provided some response to the allegations in my letter sent to you on October 1, 2021 (henceforth "my October 1 letter"). Professor Johnson, who is both the reviewer and the author of the relevant papers, responded to my letter on October 11, 2021 with additional allegations (henceforth "Johnson's October 11 letter"). Thank you for the opportunity to provide additional responses. I am submitting these additional responses with a request for your careful consideration. Please note at the outset the following:

(a) I was not provided an opportunity to respond to several allegations put forward in Johnson's October 11 letter to you; and these allegations are included in the Editors' letter.

(b) I was not informed that formal allegations would be put forward prior to an opportunity to provide evidence that disproves the accusations of substantive plagiarism that are presented as accurate in the Editors' letter. Had I known that I would not be given this opportunity, I would have included this evidence with my October 1 letter. I did not include this evidence with that letter since I believed it would be best to give as concise a response as possible. The Editors accepted Johnson's allegations without having seen this written evidence.

(c) There are serious concerns that the Editors' acceptance of Johnson's allegations was formed without proper knowledge of what is common currency in the literature on biased algorithms. The places where Johnson and my ideas overlap are places where they overlap with many others in the literature. The norm in this literature (as it is in most literatures) is not to include citations on matters that are common currency.

(d) Research misconduct is defined as "departing significantly from accepted practices of the relevant research community; and the misconduct was committed intentionally, knowingly, or recklessly; and the allegation be proved by a preponderance of evidence" (COPE 2019).[1] As noted in Reznik's seminal paper: "It is important to distinguish between misconduct and honest error or a difference of scientific opinion to prevent unnecessary and time-consuming misconduct proceedings, protect scientists from harm…" (2012, p. 1). There are serious concerns that the evidence available to the Editors at the time they accepted these allegations as credible was not sufficient for an accusation of research misconduct.

(e) There are serious concerns that the Editors' acceptance of these allegations was based on a misinterpretation of one of the documents sent by Johnson. The Editors' letter refers to Johnson's "NSF grant." The relevant document is a rough draft of an application for a grant.[2] I had never seen a draft of this grant application, or even heard of it, until it was sent to me by the Editors (I could not possibly have taken from it). There is no evidence to suggest I had seen it prior to September 27, 2021.

[1] See https://publicationethics.org/files/cope_webinar_allegations_of_misconduct_2019.pdf

[2] Here is a quote from the first page of this document: "PLACEHOLDER FOR AUTHOR ATTRIBUTIONS/DATES/ETC. – not sure how this should be written. Can you draft, Abby? Should include Hiram, Abby, Michelle, me, Gabbrielle Johnson. I also got some assistance from Branwen Williams."

I firmly deny all allegations of research misconduct. In Section 1, I provide critical background that undermines all allegations concerning substantive plagiarism. In Section 2, I provide specific rebuttals to those allegations, including the three to which I was not given a chance to respond before the Editors contacted Rutgers University. In Section 3, I respond to accusation of textual plagiarism regarding (1) seven sentences in three online sources (this was an honest mistake), (2) three passages in Johnson's published work of which the Editors note "some of these instances are less clear than the ones taken from the online sources and the NSF grant", and (3) the draft of Johnson's NSF application. All referenced documents can be found here:
https://www.dropbox.com/sh/dk4ojr7f09fifjq/AACgtGaBQl6U_auIlinODk9Xa?dl=0

These are the key points in this letter:

(1) The core ideas, motivation, and structure of my 2018 *NYT* submission, my 1500-word 2020 *New Statesman* publication, and my *Philosophical Studies* submission are the same. All allegations about overlap of ideas concern ideas that are in my 2018 *NYT* submission. I read a paper of Johnson's for the first time on December 30, 2019.[3] So these ideas in my paper do not stem from Johnson's work. The material in the section on fine-tuning bias and the sections on the different stages and levels at which biases can occur is new. The allegations about overlap of ideas does not concern the new material.

(2) Anyone who knows the literature will see that the places where Johnson and my ideas overlap are places where they overlap with many others in the literature on biased algorithms. The ideas are common currency in the literature and I do not know of any article that cites authors regarding the relevant ideas.

(3) In my *Philosophical Studies* submission, I discuss, quote, and cite Johnson's work more than the work of anyone else. I contrast my view from hers on p. 6 and p. 9. I showcase her work on p. 11. Johnson's view is the only view from which I contrast my view. Johnson is the only author I quote in the submission. Johnson is the only author who appears twice in the list of references.

(4) My idea of bottom-up biases is different from Johnson's idea of truly implicit biases as is evidenced in my *Philosophical Studies* submission. I clarify this distinction further in my October 1 letter to the Editors. In her October 11 letter, Johnson notes of the clarification: "I agree that this definition makes the theoretical analyses different." There is no change in my view between my submission and my October 1 letter.

(5) I saw the draft of Johnson's NSF grant application for the first time when it was sent to me by the Editors on September 27, 2021. I could not possibly have taken from it (textual overlap 11-15 and 17 mentioned in Editors' letter). September 27, 2021 was also the first time that I heard of this grant application. The relevant sentences are in a 5-page unpublished document that Johnson and I wrote in the context of our Lebowitz prize application. The sole purpose of this document was to provide material about biased algorithms to the three philosophers who wrote letters of recommendations in support of our Lebowitz prize application.

(6) Re textual overlap 7-9: There is no doubt overlap between how Johnson and I present the relevant basic facts about machine learning—as there is between how she, I, and many other authors present those same facts.

(7) The overlap with 3 online sources (unrelated to Johnson's work) is an honest mistake which I deeply regret (textual overlap 1-6, 10a, and 16). Further, I regret that where I showcase Johnson's work, I made the mistake of staying too close to her formulations (textual overlap 10b). I apologize for the mistakes I have made. None of these mistakes were made intentionally, knowingly, or recklessly. I apologize also for any ways in which I contributed to misunderstandings between Johnson and myself.

[3] See Johnson's October 11, 2021 letter to the Editors for confirmation that I read a paper of Johnson's for the first time on December 30, 2019.

To facilitate assessment, I include this timeline:

Jan. 27, 2018	My first talk on biased algorithms in AI and human vision at *Brooklyn Public Library*.
March 7, 2018	My paper "Biased Algorithms and AI and Human Vision" submitted to the *New York Times*.
May 30, 2018	Short version of my paper submitted to the *New York Times*.
Fall 2018	I teach a 400-level undergraduate seminar at Rutgers entitled "The Neural Basis of Cognition"
Spring 2019	I teach a graduate seminar at Rutgers entitled "Information Processing, Neural Networks, and AI"
Feb. 2, 2019	My second talk on biased algorithms in AI and human vision at *Brooklyn Public Library*
March 2, 2019	My paper "Biased Algorithms in AI and Human Vision" submitted to *New Statesman* and accepted for publication.
Dec. 30, 2019	I read a paper by Johnson for the first time.
April 2020	My paper "Biased Algorithms in AI and Human Vision" published in *New Statesman* under the title "How biased algorithms perpetuate inequality" (title selected by the editor).
June 2020	Johnson's paper "Algorithmic Bias" published online at *Synthese*.
Oct. 2020	Johnson's paper "The Structure of Bias" published in *Mind*.
Jan. 5, 2021	Johnson and I jointly apply for the Lebowitz prize.
Aug. 1, 2021	Submission deadline of special issue of *Philosophical Studies* to which I was invited to submit.
Sept. 4, 2021	Johnson contacts *Philosophical Studies* and myself with concerns about my paper.
Sept. 27, 2021	I receive a letter from the Editors with Johnson's allegations.
Oct. 1, 2021	I send my response to the Editors.
Oct. 11, 2021	Johnson sends her response to the Editors. This letter includes new allegations.
Nov. 5, 2021	The Editors send their letter to Rutgers.

1. Relevant background that undermines all allegations concerning plagiarism of ideas

I have been reading the literature on biased algorithms widely since 2017. In 2017, I was invited to give a 20-minute talk at the *Brooklyn Public Library* on a socially relevant topic. Since reading Cathy O'Neil's 2016 award-winning book *Weapons of Math Destruction* that brought biased algorithms to national attention, I had been interested in bias in AI and how the same kind of biases might also occur in human vision. So I decided to give the talk on biased algorithms in AI and human vision.

I gave the talk on January 27, 2018. I submitted the written version of that talk to be considered for publication in the *New York Times* on March 7, 2018 and a shorter version on May 30, 2018 (emails to and from *NYT* and the articles submitted attached).[4] The submission as well as the talk include all the ideas that I have allegedly taken from Johnson. I make the distinction between top-down and bottom-up biases in the very same way as I do on p. 1 of my *Philosophical Studies* submission (though at the time I call them algorithmic biases rather than bottom-up biases):

"Biases in human vision are typically discussed in the framework of top-down biases. Top-down biases stem from the effects that our concepts and beliefs have on our perception. If we have racist and sexist beliefs, this is likely to affect how we

[4] The NYT wanted to publish it, but then I dropped the ball. So this piece never got published. The NYT did not respond to my March 7, 2018 email. I learned that submissions over 2000 words are ignored. So I sent a shorter version on May 30, 2018. In her response, Johnson questions that I gave a talk on biased algorithms in January 2018. I gave a talk on the topic of biased algorithms in AI and human vision on both January 2018 and February 2019 though only the title of the second talk was "Biased Algorithms in AI and Human Vision".

perceive people. Such top-down biases exist, but they need to be distinguished from algorithmic biases. Algorithmic biases stem not from the effects of our beliefs on perception, but rather from how our visual system functions at the lowest level: the most primitive levels of visual processing." (*NYT* submission May 30, p. 1; *NYT* submission March 7, p. 2)

In Fall 2018, I taught an undergraduate class in which we discussed biased algorithms. In Spring 2019, I taught a graduate seminar entitled "Information Processing, Neural Networks, and AI", in which these issues were discussed more extensively. Since early 2018, I have been meeting and emailing with researchers to discuss biased algorithms in AI and human vision, in particular Per Anderson (Google) and Professor Tina Eliassi-Rad (previously at Rutgers, now at NorthEastern University). (Sample email about biases in AI and human vision attached.)

In February 2019, I gave a more mature version of my 2018 talk at the Brooklyn Public Library. This led to several invitations, including my piece published in *New Statesman* (submitted on March 2, 2019; email attached), a talk in an invited symposium at the *2021 APA Pacific Division Meeting*, and an invitation to publish a written version of that talk in a special issue of *Philosophical Studies* (along with other papers presented at the 2020 and 2021 *APA Pacific Division Meetings*).

The paper I submitted to *Philosophical Studies* is an expansion of my March 2018 *NYT* submission and my March 2019 *New Statesman* submission. The material in the section on fine-tuning bias and the sections on the different stages and levels at which biases can occur is new. None of the allegations about overlap of ideas concern the new material.

Johnson. As the above shows, I have been steeped in the literature on biased algorithms long before December 30, 2019, when I first read one of Johnson's papers. I admire Johnson's work. She is the only author who appears twice in the list of references in my submission (which includes only 14 references). She is the only author who I quote in the submission. She is the only author whose view I distinguish from mine (p. 6 and p. 9). She is the only author whose work I showcase (p. 11).

While I am a fan of Johnson's work, her work in no way influenced the core ideas in my submission, their motivation, or presentation. Evidence for this is that the core ideas, motivation, and structure of my 2018 *NYT* submission, my 1500-word *New Statesman* publication, and my *Philosophical Studies* submission are the same. All allegations about overlap of ideas concern ideas that are in my 2018 *NYT* submission. I am grateful that Johnson acknowledges in her October 11 letter to the Editors that these ideas in my submission and my *New Statesman* article are the same. Further, where Johnson and my ideas overlap, they overlap with many others in the literature. They are common currency in the literature.

As mentioned, I cite, quote, and discuss Johnson's work. There are many authors whose work on low-level biases I could have (and perhaps should have) discussed more—including Johnson's. I am known to over-cite and my submission is an extreme counterexample to my typical citation practice. I neither discuss, quote, or even reference Cathy O'Neill's 2016 book, Kearns and Roth's 2019 book, and Tina Eliassi-Rad's work, even though I read and thought about their work long before I read any of Johnson's work. These are just a few omissions in my submission. I did not take ideas from any of these authors and where there is overlap of ideas, they are common currency in the literature. The norm in this literature is not to cite ideas that are common currency. However, by the standards of my own citation practices, my submission under-cites the literature and for this I am deeply sorry.

How did this happen? The paper was initially written in the style of a breezy public philosophy piece and so with minimal citations. To meet the August 1, 2021 deadline for the special issue of *Philosophical Studies* to which I was invited to contribute, I had to scramble to get the paper from its original state as a breezy public philosophy piece into the shape of a journal submission. I did so while single parenting a child who was deeply struggling with the consequences of the pandemic. This is not an excuse, but it might explain why this submission is in every respect far below the level of any other of my submissions and far below my published work. In hindsight I realize that I should have rejected the invitation to submit the paper to the special issue of *Philosophical Studies*. I have learned my lesson.

I am extremely regretful about the honest mistakes I have made and I apologize. Integrity and honesty are at the very core of my values and in this instance, I failed to live up to my values. I am especially regretful about any harm that my actions have caused Johnson. I work hard to support junior women in the profession and so I am dismayed by Johnson's allegations. I am a big supporter of Johnson: I have given her comments on her papers, advice on how to respond to reviewers, invited her to my home, invited her to jointly apply for the Lebowitz prize etc. In my email asking for letters of recommendations from Ram Neta (UNC), Chris Hill (Brown), and Nico Orlandi (UCSC) in support of our Lebowitz prize application, I write "Gabbrielle is doing extremely interesting and cutting edge work on biased algorithms with deep knowledge of both the philosophical and technical aspects of the issue" (this and further emails attached). In my October 1 letter to the Editors, I include paragraphs that I added to a revised version of the paper in which I discuss Johnson's work in more detail and I cite Johnson on everything on which she requests to be cited. I apologize for the mistakes I have made and any ways in which I might have contributed to misunderstandings between Johnson and myself.

2. Response to allegations of plagiarism of ideas (for details see Appendix I)

I am confident that anyone who knows the literature will see that the places where Johnson and my ideas overlap are places where they overlap with many others in the literature. I quote the allegations from the Editors' letter (henceforth "relevant ideas"):

1. They both aim to explain the presence of bias in systems that might seem as though they couldn't be biased.
2. They both do so by positing that there are two kinds of bias: an often discussed and reasonably well understood kind of high-level bias, and a poorly understood and less discussed low-level bias.
3. They both use the same cases to illustrate the high-level bias they posit.
4. They both use the same cases to illustrate the low-level bias they posit.
5. They both claim that the relevant exemplars of both high-level and low-level biases can be found in both humans and algorithms.
6. Johnson's work calls the low-level bias a "truly implicit" bias. What makes a truly implicit bias a "low level" bias is that it is not represented by the biased system (either human or machine). Schellenberg's work calls the low-level bias she posits a "bottom-up" bias.

i. Relevant ideas 1, 3, 4. These three allegations in the Editor's letter were not in the original complaint I received from the Editors on September 27, 2021. They are among several new allegations made in Johnson's October 11 letter. Since they were not included in the original complaint, I did not have a chance to respond to these three allegations before the Editors contacted Rutgers University on November 5, 2021. To my knowledge, this constitutes a breach of protocol.

ii. Relevant ideas 1-5. (a) The relevant ideas 1-5 and my idea of bottom-up biases (I call them algorithmic biases at the time) were in my January 2018 talk, which I submitted to the *New York Times* on March 7, 2018, and a version of which was ultimately published in *New Statesman*. So I have written evidence that these ideas were in my written work for over two years before any of Johnson's work was available online and almost two years, before Johnson sent me her work. I mention, but did not include this written evidence in my response. The Editors accepted the allegations without having seen this written evidence.

(b) Moreover, the relevant ideas 1-5 are common currency in the literature on biased algorithms. I have not seen an article that cites anyone on 1-5, as is typical for material that is common currency. Despite not citing anyone on 1-5 herself, Johnson accuses me of "poor scholarship" for not doing so. I kept my October 1 letter short and so did not include quotes to substantiate that 1-5 are common currency in the literature—indeed I did not address 1, 3, 4 at all, since they were not part of the original complaint. I realize now that this was a mistake.

(See Appendix I for quotes showing that 1-5 are in my 2018 *NYT* submission, quotes that substantiate that 1-5 are common currency in the literature, and more detailed discussion of 1-5. See Appendix III for response to Johnson's October 11 letter.)

In her October 11 letter, Johnson writes "I understand Schellenberg's response as granting the overlap I've noted in 1-5". In response: Since 1, 3, and 4 were not in her original complaint, I did not have the chance to respond to those allegations, and so could not possibly have granted overlap. I struggle to understand how Johnson makes the leap from 1-5 being common currency to me having taken 1-5 from her. Does Johnson believe 1-5 are her ideas? Does she not know that 1-5 are common currency in the literature? Does she believe that I should cite her on these ideas even though the norm of this literature is not to provide citations on ideas that are common currency? Does she believe my reading her papers was my window into this literature? I have been working on biased algorithms since 2017, am well-versed in the literature, have taught an undergraduate class and a graduate seminar in which we covered the topic, and have given a talk in January 2018 that includes 1-5. In March 2018, I submitted a written version of that talk to the *New York Times*.

iii. Relevant idea 6. My bottom-up biases and Johnson's truly implicit biases are different. To make this fact more explicit than it already is in the submission, I added the following to my paper and quoted it in my October 1 letter:

> "It will be helpful to distinguish bottom-up biases from implicit biases. Bottom-up biases can be implicit, explicit, conscious, or unconscious and they may or may not be represented. In the very same way, top- down biases can be implicit, explicit, conscious, or unconscious, and they may or may not be represented. The difference between top-down and bottom-up biases is exclusively a difference in their *source*. As argued above, top-down biases stem from beliefs, desires, goals, fears, and other such mental states of human agents (that may or may not be propositional attitudes). By contrast, bottom-up biases stem from how recognitional systems process data at the lowest level.[5]"

In her October 11 letter to the Editors, Johnson writes about this passage:

> "I agree that this definition makes the theoretical analyses different." And "Schellenberg's newest characterization of "bottom-up" biases … does indeed make them different from my truly implicit biases. … In this passage, Schellenberg explicitly tells us that what makes a bias "lower level" for her is not anything having to do with representation, but instead is entirely the bias's (causal?) source."

I am grateful to Johnson for acknowledging this. Yet, she adds that this "appears nowhere in the paper, but only makes an appearance now in her letter to you." In response: There is no difference between my view in my submission and how I present it in my October letter. As I write in my submission:

> "To a first approximation top-down biases *stem* from the beliefs and world views of programmers, whereas bottom-up biases *stem* from incoming data and its processing at the lowest levels of AI systems" (p. 1; italics added; for more details, see pp. 2-3).

> "The argument of my paper is neutral on whether biases are representational or functional." (p. 6).

So it is clear in my submission that the distinction between top-down and bottom-up biases is what they *stem* from, that is, their *source*. That same characterization of the differences is in my 2018 *NYT* submission. I do not discuss representations in the paper other than to say that everything I say is neutral on whether biases are representational.

[5] For an argument that all implicit biases have representational content, see Mandelbaum (2015). For an argument, that at least some are non-representational, see Johnson (2020). Johnson develops a non-representational, functionalist account of a special kind of implicit biases and argues that algorithmic biases are an example of such implicit biases. On my view, such non- representational, functional implicit biases could be bottom-up, but they could equally be top-down biases. As argued above, many algorithmic biases are top-down biases. Moreover, bottom-up biases could have representational content. In short, the distinction between top-down and bottom-up biases is orthogonal to the distinction between implicit and explicit biases. It is neutral on all matters other than the source of the biases.

Since all parties seem to agree that Johnson's idea of truly implicit biases differs from my idea of bottom-up biases, I will move on to address allegations of textual overlap. However, if there is any doubt that my idea is similar to hers, I would be happy to provide more detail on this matter.

3. Response to allegations of textual plagiarism

iv. All allegations of textual overlap 1-18 listed in the Editors' letter concern sentences that present either universally accepted basic facts about machine learning or widely discussed examples of bias in AI or human vision. None are in the context of presenting my original ideas.

v. Textual overlap 1-6, 10a, and 16 mentioned in Editors' letter. I am deeply embarrassed by the textual overlap in 7 sentences with 3 online sources. This was an honest mistake for which I take full responsibility and for which I apologize profusely. (For details, see Appendix II.)

Regarding textual overlap with Johnson's work, I am grateful that the Editors' letter acknowledges that "some of these instances are less clear than the ones taken from the online sources and the NSF grant".

vi. Textual overlap 11-15 and 17. These concern the rough draft of Johnson's NSF grant application (henceforth "draft" or "Johnson's draft"). I saw this draft for the first time on September 27, 2021 when it was sent to me by the Editors. 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 passages in 11-15 and 17 are all in a 5-page unpublished document that Johnson and I jointly wrote in December 2020. In December 2020, I suggested that Johnson and I jointly apply for the Lebowitz Prize. This prize is given yearly to two philosophers who hold contrasting views on a philosophical question of public interest. The two philosophers must be (self-)nominated jointly. None of the philosophers who wrote letters of recommendations in support of our application had prior knowledge of the topic. So Johnson and I wrote a 5-page document outlining issues about biased algorithms that we sent to each of them. We did not win the prize and agreed that the material of our joint application can be used freely by each of us. If there is any disagreement about this, then I apologize profusely for my part in the poor communication that led to this misunderstanding.

I will add that 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). 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 paper. However, on careful inspection, I see that Johnson did provide some material that made it into pages 1-3. This is the material in the textual overlap 11-15 and 17. 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).

vii. Textual overlap 7-9. Most papers on biased algorithms include some version of these sentences. If one does an online search for "There are three types of machine learning programs: supervised learning, unsupervised learning, and reinforcement learning" one gets a plethora of results—some closer to my formulation; others closer to Johnson's.

The same holds for the basic fact that supervised learning includes a training phase and a test phase. Further, since supervised learning is the simplest of the three types of machine learning, almost all discussion of biased algorithms focuses on supervised learning. In the documents sent to the

editors, the terms "label" and "features" are highlighted, suggesting that those are Johnson's terms. But these are the terms used by everyone in this literature. I include a few examples in the footnote.[6]

There is no doubt overlap between how Johnson and I present these basic facts about machine learning—as there is between how she, I, and many other authors present those same facts.

Textual overlap 10b. These sentences are from a short paragraph about proxy attributes. Proxy attributes are the focus of Johnson's paper "Algorithmic Bias". They are irrelevant for the issues discussed in my paper. Nonetheless, I mention them with the aim of showcasing Johnson's work. In doing so, I stayed too close to her formulations. I apologize sincerely (as I have already done to Johnson directly). I cite her right after the relevant sentence quoted under 10b in the Editor's letter.

Please don't hesitate to get in touch if you have questions.

Sincerely,

Susanna Schellenberg

Appendix I. Discussion and quotes showing that 1-5 are common currency in the literature and were in my 2018 New York Times submission.

I structure the discussion and quotes according to the six allegations regarding overlap of ideas from the Editors' letter. I include quotes only from widely read sources, TED talks that have been seen over 1 million times, and similar such venues. If the quote is from a scientific article, it is from the abstract of that article (to show that the relevant point is central in the article). A lot of the literature on AI bias is published in publication venues that are exclusively online. So I include references to such sources. All quotes were available online before December 2019 and so before either Johnson or I had published anything on the topic.

1. **"They both aim to explain the presence of bias in systems that might seem as though they couldn't be biased."**

I did not have the opportunity to respond to this allegation prior to the Editors' letter being sent to Rutgers University. The claim that algorithms are biased even though they seem as if they shouldn't be is made in almost every article on biased algorithms. It is in the opening paragraph of both of my *NYT* submissions:

"Artificial intelligence systems and human visual systems are riddled with biases. Given how flawed humans are, it's not surprising that our vision is biased. But it may come as a surprise that AI systems are. After all, computer algorithms form the core of AI systems and computer algorithms are grounded in mathematics and operate on data. So one might expect them to be objective and just. But they aren't." (Schellenberg, March 7, 2018 *NYT* submission, p. 1)

(a) "Algorithms are opinions embedded in code. It's really different from what you think most people think of algorithms. They think algorithms are objective and true and scientific. That's a marketing trick." (O'Neill's 2017 TED Talk)

[6] Here are two examples: "In supervised learning, features are learned via labeled input…." (https://patents.justia.com/patent/10074038) and "In the training phase, the algorithm takes two parameters as input. First is the set of features, and second is the classification labels" (https://www.msystechnologies.com/blog/how-to-use-naive-bayes-for-text-classification/). Here is an example that includes more information: "Like all machine learning algorithms, supervised learning is based on training. During its training phase, the system is fed with labeled data sets, which instruct the system what output is related to each specific input value. The trained model is then presented with test data: This is data that has been labeled, but the labels have not been revealed to the algorithm. The aim of the testing data is to measure how accurately the algorithm will perform on unlabeled data." (https://searchenterpriseai.techtarget.com/definition/supervised-learning).

(b) "A promise of machine learning in health care is the avoidance of biases in diagnosis and treatment; a computer algorithm could objectively synthesize and interpret the data in the medical record. Integration of machine learning with clinical decision support tools, such as computerized alerts or diagnostic support, may offer physicians and others who provide health care targeted and timely information that can improve clinical decisions. Machine learning algorithms, however, may also be subject to biases. The biases include those related to missing data and patients not identified by algorithms, sample size and underestimation, and misclassification and measurement error. There is concern that biases and deficiencies in the data used by machine learning algorithms may contribute to socioeconomic disparities in health care." (Gianfrancesco et al 2018: abstract).

(c) "Rather than clinging to the belief that technology is impartial, engineers and developers should take steps to ensure they don't accidentally create something that is just as racist, sexist, and xenophobic as humanity has shown itself to be." (Garcia 2017)

2. "They both do so by positing that there are two kinds of bias: an often discussed and reasonably well understood kind of high-level bias, and a poorly understood and less discussed low-level bias."

It is standard to distinguish biases that stem from human cognition from biases that stem from data and its processing. Again, I have never seen any article or book be cited when this claim is made, but in hindsight I realize that I should have cited various sources just to show that this idea is common currency and to make clearer what the original contribution of my article is (namely the different kinds of bottom up biases). I make the distinction in my *NYT* submissions:

"Top-down biases stem from the effects that our concepts and beliefs have on our perception. If we have racist and sexist beliefs, this is likely to affect how we perceive people. Such top-down biases exist, but they need to be distinguished from algorithmic biases. Algorithmic biases stem not from the effects of our beliefs on perception, but rather from how our visual system functions at the lowest level: the most primitive levels of visual processing." (Schellenberg, March 7, 2018 *NYT* submission, p. 1)

Here is an example where the distinction is introduced in the very definition of algorithmic bias:

(a) "Algorithmic bias---when seemingly innocuous programming takes on the prejudices either of its creators or the data it is fed---…" (Garcia 2017)

Here are a few further examples:

(b) "Are violations of privacy and fairness the result of incompetent software developers or, worse yet, the work of evil programmers deliberately coding racism and back doors into their programs?

The answer is a resounding no. The real reasons for algorithmic misbehavior are perhaps even more disturbing than human incompetence or malfeasance (which we are at least more familiar with and have some mechanisms for addressing). Society's most influential algorithms—from Google search and Facebook's News Feed to credit scoring and health risk assessment algorithms—are generally developed by highly trained scientists and engineers who are carefully applying well-understood design principles. The problems actually lie within those very principles, most specifically those of machine learning." (Kearns and Roth 2019, p. 14)

(c) "supervised and unsupervised learning models have their respective pros and cons. Unsupervised models that cluster or do dimensional reduction can learn bias from their data set. If belonging to group A highly correlates to behavior B, the model can mix up the two. And while supervised models allow for more control over bias in data selection, that control can introduce human bias into the process." (Lynch 2018)

(d) "An important distinction to make at this point is that such bias showing up in AI isn't an automatic sign of deliberate and malicious injection of the programmers' bias into their projects. If anything, these AI programs are simply reflecting the example bias that already exists. Even if AI is trained using a vast amount of data, it can still easily pick up patterns within that lead to problems like

gender assumptions because of the range of published material that contain these linked words." (Murray 2019)

(e) "It doesn't take active prejudice to produce skewed results …. It just takes distorted data that no one notices and corrects for." (Garcia 2017)

(f) "Machine learnt systems inherit biases against protected classes, historically disparaged groups, from training data. Usually, these biases are not explicit, they rely on subtle correlations discovered by training algorithms, and are therefore difficult to detect." (Datta et al 2017, abstract)

(g) "Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with." (Barocas and Selbst 2016, abstract)

As is the case in Johnson's paper, the distinction is often made at some point along the way, rather than being the focus of the paper.

3. "They both use the same cases to illustrate the high-level bias you posit."

I did not have the opportunity to respond to this allegation prior to the Editors' letter being sent to Rutgers University. The examples Johnson lists are surprising. They are "discriminatory (racist and sexist) beliefs" and "biases programmers write in software code".

Biased beliefs are the classic case of a high-level bias. Racism and sexism are the most common forms of discrimination. If such beliefs of programmers enter an algorithm, then one key way in which they will is in virtue of the programmers writing code. Almost every article on biased algorithms mentions these issues. Here is how I make the point in my 2018 NYT submission:

"Top-down biases stem from our concepts, beliefs, and worldview filtering into how we perceive the world. If we have racist and sexist beliefs, this is likely to affect how we perceive our environment." (Schellenberg, March 7, 2018 *NYT* submission, p. 2).

[NB: In my *NYT* submission, I do not explicitly state that one way in which the biases of programmers enter an algorithm is via the programmer writing code. However, this is implied.]

I include only one quote to show that this is common currency in the literature:

"The biases of humans that program and apply the algorithm can translate into the algorithm, and sometimes stereotypes and negative associations can be codified in and amplified by the algorithm. Algorithms, after all, are built, trained, and implemented by people that, like everyone else, have prior beliefs and goals determined by social factors. Researchers have known for a long time that people's prior beliefs translate into the algorithms that they generate. Common behavioral biases that could turn into a model's assumptions are, for example, reporting bias, selection bias, and availability bias. These could be translated, for example, into faulty labeling, faulty personalization that leads to filter bubbles or tunnel vision, incorrectly assuming causation, or one side of the matching incorrectly given an advantage over the other side. Facial recognition machine learning software, for example, have been shown to be affected by the demographics of the people who design them. This issue is exacerbated if programmers are disproportionately male, white, and heterosexual. … In sum, there is always, at some level, a human decision-maker that impacts the process. Biases in an algorithmic process often exist because human biases were translated into the system." (Cofone 2019, p. 1401).

Cofone then distinguishes such biases from data-driven biases "see in particular the section entitled "It's All in the Data". So this is yet another example of (2) above.

4. "They both use the same cases to illustrate the low-level bias you posit."

Again, I did not have the opportunity to respond to this allegation prior to the Editors' letter being sent to Rutgers University. All papers on biased algorithms of which I am aware include long lists of examples of biased algorithms. As is typical in this literature, Johnson and I each list lots of examples

of biases. There are lots of biases mentioned by me, not mentioned by Johnson and vice versa. Where there is overlap, they are biases that are discussed by almost everyone in the literature.

With the exception of biases that come from poor labeling/classification practices, all biases on Johnson's list of overlapping cases are mentioned in my 2018 *NYT* submissions. Here is her list (in italics) along with a quote of the sentence in which the relevant example first comes up in my March 7, 2018 NYT submission:

speech recognition: "Google's speech recognition algorithm is a good example." (Schellenberg, March 7, 2018 *NYT* submission, p. 1) [Contrary to what Johnson claims, I don't discuss natural language processing in any of my papers beyond speech recognition software. So I don't include discussion of that here or below.]

hiring algorithms: "In various companies around the country, computer algorithms are used to filter out promising job applications." (Schellenberg, March 7, 2018 *NYT* submission, p. 4)

Google ad software: "Other examples of feature-linking bias are Google's online advertising algorithm." (Schellenberg, March 7, 2018 *NYT* submission, p. 4)

facial recognition algorithms: "A Nikon camera with built-in facial recognition software used to misrepresent Asian faces as blinking, asking the user, "Did someone blink" when the subject of the photo was Asian." (Schellenberg, March 7, 2018 *NYT* submission, p. 1)

recidivism-risk software COMPAS: "One is the Northpointe "recidivism risk" software Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS." (Schellenberg, March 7, 2018 *NYT* submission, p. 4)

training data bias (Johnson calls it "Biases that come from unrepresentative data", the more standard label is "training data bias" or "training sample bias".): "One source of bias is the data on which the algorithm is trained. We can call this training-data bias. … the software is customized to its operational environment via training data. It turns out that Google's speech recognition algorithm was trained on unbalanced data. The training data consisted disproportionately of male voices. As a consequence, the algorithm is better at decoding voices within the range of the typical frequency of male voices." (Schellenberg, March 7, 2018 *NYT* submission, p. 1)

biases that come from injustice in the environment: "Training data bias is a problem since algorithms that operate on biased training data lead to racist or sexist outcomes. By repeating our past practices, algorithms not only automate the status quo and perpetrate bias and injustice, they can amplify the biases and injustices of our society. Just to be clear: repeating our past practices would be great if we lived in a perfect world. In our imperfect world, algorithms trained on biased data generate harmful feedback loops and reinforce human prejudices." (Schellenberg, March 7, 2018 *NYT* submission, p. 1)

light-comes-from-above bias: "Let's first talk about the beneficial biases. Our perceptual system assumes that light comes from above. This is a beneficial bias, since light in fact typically comes from above." (Schellenberg, March 7, 2018 *NYT* submission, p. 2) [Contrary to what Johnson claims, I don't discuss distance calculation or visual perceptual heuristics more generally in my submission. So I don't include discussion of that here or below.]

As to evidence that these examples are common currency in the literature, with exception of the light-comes-from-above bias, the following articles and books discusses every one of the biases on Johnson's list: O'Neill 2016, Penchikala 2018, Groen 2018, Martin 2019, Cofone 2019, Kearns and Roth 2019, Cowgill 2019 (and a long list of further articles and books). There are many more that mention all but two of the examples of biases in the list above.

In O'Neil's 2017 TED talk, she discusses racist and sexist hiring algorithms, the COMPAS recidivism risk software (Johnson calls it risk assessment software), biases that come from injustice in the environment, and training sample biases.

If one searches for the following on Google: "voice recognition, hiring, Google ads, facial recognition, recidivism-risk compas, training sample bias, labeling", one gets 23'300 results.

5. They both claim that the relevant exemplars of both high-level and low-level biases can be found in both humans and algorithms.

It is hard to find an article on algorithmic bias that doesn't talk about the biases that occur in AI also occurring in humans. It has been observed by many that the same kind of bias can occur in both AI and the human mind. This is not surprising. After all, the cutting edge of AI and neuroscience intersect: AI looks to neuroscience, and in particular brain scientists, to learn how best to create machines that

exhibit intelligence; neursoscience looks to AI to model the brain. So similarities between artificial and human intelligence are unlikely to be missed. The interesting question is what the nature of theses biases are.

The observation that the same kind of biases occur in humans and AI is in my 2018 NYT submission. Indeed, my January 2018 talk and my 2018 *NYT* submissions are about biases in humans and AI to roughly the same degree. See also my April 2020 *New Statesman* article. I should note that I don't explicitly mention top-down biases in AI in those papers, but since top-down biases stem from programmers coding algorithms, the fact that I distinguish them with regards to humans implies that the humans programming algorithms contribute to biases in AI. Here are a few examples from the literature:

(a) "Human biases are well-documented, from implicit association tests that demonstrate biases we may not even be aware of, to field experiments that demonstrate how much these biases can affect outcomes. Over the past few years, society has started to wrestle with just how much these human biases can make their way into artificial intelligence systems — with harmful results." (Manyika 2019)
(b) "An algorithm is a very general concept - it's something that we do actually in our heads every day. To build an algorithm we need only two things, essentially: a historical data-set and a definition of success." (From O'Neill's 2 ½ min summary of her 2016 book.)
(c) "everyone uses algorithms. They just don't formalize them in written code. Let me give you an example. I use an algorithm every day to make a meal for my family." (From O'Neill's 2017 TED Talk.)

Appendix II. Details on (v): seven sentences in which there is textual overlap with three online sources (unrelated to Johnson).

My only explanation for how the five sentences that overlap with the Wikipedia article on cross-race bias could have happened is that my procedure for using my notes was impacted when putting the finishing touches on this paper.[7] [Details in footnote] Like many academic parents, I was working under unusual circumstances this past year. I was doing so as a single parent of a young child who was deeply struggling with the consequences of the pandemic. This is not an excuse and I sincerely apologies for this honest mistake. I should note that these five sentences detail widely known examples of cross-race bias. They do not express original ideas.

The two sentences that overlap with the two other online sources were in the notes I took while auditing a graduate seminar on machine learning in Spring 2019 (in the context of my *Mellon New Directions Fellowship*). I do not recall reading either of those two online sources, but I must have read them at the time. I am deeply sorry for this honest mistake. The two sentences describe basic facts about machine learning on which all parties agree. So here again the issue is a matter of presentation of universally accepted ideas.

Appendix III. Response to Johnson's October 11 Letter to the Editors (henceforth "reply")

In her reply, Johnson does not address the following facts addressed in my October 1 letter:

(a) The relevant ideas that she claims I took from her are common currency in the literature.

[7] When I started working on bottom-up biases in 2017, I created a document in which I collected examples of such biases in AI and human perception—examples that I came across as I was reading newspapers, scientific articles, and apparently Wikipedia articles. The examples on p. 4 of my submission were the first I included in this document. Shortly after, whenever adding examples, I described them in my own words. But it appears that in the very first instance I added the examples by copying the material with only minimal changes from the text in which I had read about them.

(b) I have been working on this topic since 2017 and so long before I had ever heard of Johnson or read any of her work and more than 2 years before she had any publications. So I do not have the relevant ideas (which are common currency) from her, but due to my knowledge of the literature.

(c) I have never seen the draft of her NSF grant application until it was sent to me by the Editors on September 27, 2021 and so could not possible have taken material from it. (Johnson does not mention the draft of her NSF application at all.)

(d) We jointly applied for the Lebowitz prize in the process of which we jointly wrote a 5-page document. The passages that I allegedly took from the draft of her NSF application (which I have never seen) are all in this 5-page document.

(e) In my October 1 response, I include paragraphs from the revised version of my submission in which I cite Johnson on all matters on which she requests to be cited.

Instead of addressing any of these matters, she adds multiple allegations that were not in her original complaint.

I do not know whether Johnson was unaware that I have been working on the topic of biased algorithms since 2017. If that is the case, then this explains several of the misunderstandings that led to the current situation. I will address just four of Johnson's points:

I. Johnson writes "I would expect the bulk of her paper to be dedicated to either arguing that her distinction is preferable to mine or to making much clearer what the distinction really is. Instead the bulk of her paper is dedicated to elaborating on 1-5" and later "I do not believe I have intellectual ownership of the idea that some biases are data driven. [Here Johnson acknowledging that (2) above is common currency] …The worry is about the commonalities in philosophical questions, motivating cases, and broad strategic responses. Either these are sufficient to establish an interesting philosophical contribution for both Schellenberg and me, or these ideas are commonplace and neither her paper nor mine should be published. Assuming she thinks her paper should be published, she should cite my philosophical contributions in this area."

In response: In these passages and elsewhere in her letter, Johnson insists that what is interesting about my paper is 1-5, that is, where there my ideas overlap with hers and with many others in the literature. This is baffling. A quick glance at my paper reveals that it is not dedicated to elaborating on 1-5, but rather dedicated to distinguishing the many different levels and stages at which bottom-up biases occur as well as different kinds of such biases. That is what the philosophical contribution of my paper is and Johnson's concerns about overlap of ideas are not about these parts of my paper. I add that where there is overlap between the two of and many others in the literature is not what is interesting about Johnson's paper either—and her papers are very interesting.

As to whether the bulk of my paper should have been dedicated to Johnson's view, as mentioned earlier: (1) In my paper, I quote only one person: Johnson (p. 9 fn 7). (2) Only one person appears in the very short list of references twice: again Johnson. (3) I distinguish my view from only one person: again Johnson (p. 6 and p. 9). (4) I showcase the work of only one person: again Johnson (p. 11). The focus of the paper is not to discuss existing literature, but rather to discuss different kinds of bottom-up biases and different stages at which such biases can occur.

II. Johnson writes about cases of high-level and low-level biases (3 and 4 above): "we both group these cases in exactly the same way". In response: Everyone in the literature groups these in the same way. There is no other conceivable way to group them.

III. Johnson's work on biased algorithms is focused on whether some biases are non-representational. Despite it being explicit in my submission that "The argument of my paper is neutral on whether biases are representational or functional" (p. 6) and despite making this even more explicit in my October 11 letter, Johnson continues to insist that my view in some way has something to do with representations.

I don't know what to do about this other than to repeat that everything I say in my paper is neutral on issues having to do with representations.

IV. Further she alleges that there is "substantive structural overlap of setting up the philosophical investigation". In response: This is simply not true. Moreover, the setup of my paper is the same as in my 2018 *NYT* submission. The focus of Johnson's paper "The Structure of Bias" is not algorithmic bias. There is no mention of AI, algorithms, machine learning, algorithmic bias, or any related terms in her abstract or the 3-page introduction of her paper. In fact, there is no mention of AI in her paper and only one mention of algorithmic bias. Her paper is about biases in general and implicit biases in particular. She approaches the topic in the context of questions about social cognition. When biases in machine learning are discussed, they are mentioned as an example of her notion of truly implicit biases. So this must be where she thinks there is overlap in our ideas. Truly implicit biases are "biases that influence an individual's beliefs about or actions toward other people, but are nevertheless nowhere represented in that individual's cognitive repertoire. I call this type of bias *truly implicit bias*, and it is a counterexample to representationalism." (*The Structure of Bias*, p. 1194f). The focus of Johnson's paper "Algorithmic Bias" is on proxy attributes. I discuss those in only one paragraph and there only to showcase Johnson's work.

References

Bajorek, J. 2019. "Voice Recognition Still Has Significant Race and Gender Biases". *Harvard Business Review*.

Barocas, S. and Selbst, A. 2016. "Big Data's Disparate Impact". *California Law Review*. DOI: http://dx.doi.org/10.15779/Z38BG31

Cofone, I. 2019. "Algorithmic Discrimination is an Information Problem." *Hastings Law Journal* 70: 1391-1444. https://repository.uchastings.edu/cgi/viewcontent.cgi?article=3867&context=hastings_law_journal

Cowgill, B. et al. 2019. "Economics, Fairness, and Algorithmic Bias" *The Journal of Economic Perspectives*.

DeBrusk, C. 2018. "The Risk of Machine-Learning Bias". *MITSloan* https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/

Eliassi-Rad, T. 2018. "Just Machine Learning". https://www.usf.edu/business/documents/analytics-forum/2018/slides-eliassi-rad-tina.pdf

Garcia, M. 2017. "How to Keep Your AI From Turning Into a Racist Monster" *Wired*. https://www.wired.com/2017/02/keep-ai-turning-racist-monster/

Gianfrancesco, M. et al 2018. "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data". *JAMA Intern Med* 178: 1544-1547. doi:10.1001/jamainternmed.2018.3763

Groen, D. 2018. "How we made AI as racist and sexist as humans". *The Walrus*. https://thewalrus.ca/how-we-made-ai-as-racist-and-sexist-as-humans/

Hao, K 2019. "This is how AI bias really happens—and why it's so hard to fix". *MIT Technology Review*. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Kearns, M. and Roth, A. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press.

Lynch, V. 2018. "Three ways to avoid bias in machine learning". *Techcrunch*.

Martin, K. 2019. "Ethical Implications and Accountability of Algorithms". *Journal of Business Ethics* 160: 835-850.

Manyika, J. et al (2019). "What Do We Do About the Biases in AI?" *Harvard Business Review*. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Murray, J. 2019. "Racist Data? Human Bias is Infecting AI Development" *Towards Data Science*. https://towardsdatascience.com/racist-data-human-bias-is-infecting-ai-development-8110c1ec50c

O'Neill, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown

O'Neill, C. 2017. "The era of blind faith in big data must end". TED Talk https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end/up-next?language=en

Penchikala, S. 2018. "Analyzing and Preventing Unconscious Bias in Machine Learning". *InfoQ*

Raso, F. et al. 2018. "Artificial Intelligence and Human Rights" https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf