RUTGERS
School of Arts and Sciences

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

Web: http://susannaschellenberg.org
Email: susanna.schellenberg@rutgers.edu
Phone: 848-932-9861

October 1, 2021

Re: Submission PHIL-D-21-00991 to *Philosophical Studies*, Special Issue: Pacific APA 2020 & 2021

Dear Professor Davis and Professor Lackey,

Thank you for the opportunity to respond to the accusations that have been leveled against me. On September 4, 2021, Gabrielle Johnson emailed me directly about some of these concerns and I responded. However, she did not reply to my response, choosing instead to elevate this to the Editors in Chief. It is unfortunate that this is taking up everyone's time.

Short response: I strongly disagree with the accusations. The allegations that "versions of Schellenberg's theses (1) and (2) play a central role in Gabbrielle M. Johnson's work" has no merit. I reject all accusations of textual plagiarism as well. Some of the sentences under discussion have their source in a 5-page document that Johnson and I wrote in January 2021 in the context of a joint application (e.g. 5 below). We agreed that the material in this document could be used freely by either of us. One sentence is due to showcasing Johnson's work while unfortunately staying too close to her formulations (4 below). Others are due to honest mistakes for which I am deeply sorry (7 and 8 below). My paper was written while engaging in pandemic parenting as a single mother. That is not an excuse, but may provide some explanation. All accusations of textual plagiarism concern sentences that present either universally accepted basic facts about machine learning or widely discussed examples of bias in AI and human vision. None are in the context of presenting my original ideas. In short, I reject all accusations. Nonetheless, in the spirit of collegiality and in the hopes that this will accelerate finding a positive path forward, I have made changes in my paper in response to every allegation except those discussed under 6 (b).

Detailed response: In what follows, I respond to each allegation one by one. My response has four parts. In Part 1, I briefly outline critical background information. In Part 2, I respond to the allegation that "versions of Schellenberg's theses (1) and (2) play a central role in Gabbrielle M. Johnson's work". In Part 3, I respond to the accusations of textual plagiarism. In Part 4, I mention some observations about concerns Johnson has expressed to me over the years that provide more context for these allegations.

I do not want to overwhelm you with material in this response. I can support all statements in this response with additional documentation should the need arise. I include footnotes with links to some documentation available online. Before delving into the details, I should mention that I have been writing philosophy for almost 30 years and this is the first time that I have been accused of plagiarism. I have never heard even the slightest rumor of me taking ideas from others.

1. Background information

I first met Johnson in November 2019. She had defended her dissertation a few months earlier and sent me her paper "The Structure of Bias", suggesting we meet to discuss it. This was the first time that I read her work. I believe she contacted me since I have discussed my idea of bottom-up biases and the way such biases affect AI and human perception extensively with her PhD supervisor at multiple conferences as far back as 2017 (i.e. long before I had heard of Johnson or had read any of her work). Since meeting her, I have been supporting Johnson in multiple ways, giving her feedback on drafts of her papers, suggesting how to respond to referee reports, sharing my ideas on biased algorithms with her, and discussing hers.

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 we wrote a 5-page

document outlining issues about biased algorithms that we sent to each of them. This is relevant since our work on this 5-page document led to some of the problems of textual overlap.

2. Response to allegations that two of the three primary theses of my paper are plagiarized from work by Johnson.

1. Johnson alleges that "Schellenberg's distinction in (1) [i.e. the distinction between top-down biases and bottom-up biases] is substantially similar to Johnson's distinction drawn out in her 'Structure of Bias' paper between biases that occur in the form of fully represented, propositional attitudes (like belief), and biases that implicitly emerge from innocuous computational rules operating on a particular dataset."

In response: First, it is worth noting that the focus of Johnson's paper 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.

My bottom-up biases and Johnson's truly implicit biases are very different. 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). Even in her notion of truly implicit biases, her focus is on the human mind. She develops a non-representationalist, functionalist account of this special kind of implicit bias. Everything in my paper is neutral on representationalism, functionalism, and whether biases are implicit, explicit, conscious, or unconscious. I have added the following paragraph (and footnote) to my paper to make this explicit[1]:

"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."

"Footnote: 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."

In the unlikely case that Johnson believes she has intellectual ownership of the idea that some biases are data driven, I include this on the first page of my paper: "The idea that some biases are data driven is not new (see O'Neill 2016 and Kearns and Roth 2019). The question is what the nature of such biases are and how they relate to other biases." The idea that at least some algorithmic biases are data driven is common ground in the literature. Cathy O'Neill's book *Weapons of Math Destruction* brought the issue to national attention.

2. Johnson alleges that "Schellenberg's second claim (2) [i.e. that bottom-up biases occur not only AI but equally in cognitive and perceptual systems] is the main thesis that Johnson takes up in her "Algorithmic Bias" paper: that some data-driven biases occur just as well in artificial and natural cognitive systems".

---

[1] All the information in this paragraph was already in the paper. While I do not like repeating myself, perhaps a concise summary will be helpful to readers other than Johnson. The footnote adds new material.

In response: This accusation only holds if the first did. There is no merit to the first accusation and so the second accusation falls flat as well.

But for the sake of argument, let's ignore that her truly implicit biases and my bottom-up biases are very different and consider a different accusation that might have traction. This accusation would be that Johnson and I both argue that biases (whatever the nature of those biases might be) can occur both in AI and human perception/cognition.

In response to this different accusation (an accusation not in fact made by Johnson): The idea that the same kind of bias (whatever the precise nature of that bias might be) can occur in both AI and the human mind is not original to either Johnson's or my work. It is discussed extensively by a range of authors. In fact, I struggle to think of authors working on algorithmic bias who do not make this claim. In order to put to rest all concerns, I added the following to the revised version of my paper:

"It has been observed by many that the same kind of bias can occur in both AI and the human mind.[2] 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. I highlight that bottom-up biases occur in both natural and artificial recognitional systems to substantiate that there are bottom-up biases in any system in which there is a mismatch of processing power and data to be processed. So, even if one has no interest in AI, one should take interest in bottom-up biases."

***

Now let's assume for the sake of argument that the accusation that "versions of Schellenberg's theses (1) and (2) play a central role in Gabbrielle M. Johnson's work" had merit (it does not). I published my idea about bottom-up biases and how they occur in AI and human perception in April 2020 in a 1000-word, breezy public philosophy article in *New Statesman*.[3] Johnson's two papers were published in June 2020 and October 2020 respectively. I was invited to publish that piece on the basis of a talk at the *Brooklyn Public Library* in February 2018. The *New Statesman* piece is an abridged transcript of that talk. (Since the editor of *New Statesman* asked that I leave out all technical terms and explain the issue via examples, I do not use the term "bottom-up biases" in the written version, nor do I develop the idea in any detail.) I hasten to add that I do not expect Johnson to cite my work: her ideas are different from mine.

3. Response to the accusations of textual plagiarism

I will start with the text comparisons between various articles of Johnson's and mine, summarized in the document "Text Comparison" that I was sent by you on September 27, 2021. To facilitate assessing whether there is merit to the accusations that have been leveled, I address them in the order that the relevant text occurs in my paper and quote the relevant text as I address each allegation one by one. Unless noted otherwise, all page numbers refer to my paper.

3. "There are three basic types of machine learning programs:
   - supervised learning
   - unsupervised learning
   - reinforcement learning …" (p. 6-7)

Johnson complains that there is overlap in how I present the basics of machine learning with how she does. One of her papers includes the sentence: "Machine learning programs come in three basic forms: supervised learning, unsupervised learning, and reinforcement learning."

In response: First, there is nothing original in either my or her ideas in the relevant section. The issue here is simply presentation of universally agreed upon basic facts about machine learning. Second, most papers on biased algorithms include some version of the sentences above. If one does an online search for "There

[2] See O'Neill (2016), Munton (2019a, 2019b), and Johnson (2020b) among many others.

[3] The article can be found here: https://www.newstatesman.com/science-tech/2020/04/how-biased-algorithms-perpetuate-inequality

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. 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.[4]

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. In the hopes that this will facilitate a speedy resolution of this matter, I have rewritten this section eliminating textual overlap with Johnson (while still using the terms employed universally in this literature).

4. "It is important to note that the features linked may be proxy attributes, that is, seemingly innocuous attributes that correlate with socially sensitive attributes in the environment such that the former can serve (either purposely or accidentally) as proxies for the socially-sensitive attributes…" (p. 11)

In response: 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). Since proxy attributes do not matter for my paper, I deleted the entire paragraph. This is unfortunate since I like elevating the work of junior scholars. While I will continue to support Johnson, I believe that going forward it is best to keep a safe distance from the topics that are close to her heart.

5. "An example of a fine-tuning bias that is discriminatory due to being too fine-tuned for a marginalized group is the predictive policing algorithm PredPol…" (p. 13)

In response: The PredPol example was in the 5-page document we wrote in January 2021 in the context of our joint application. We did not win the prize and agreed that the material of our joint application can be used freely by each of us. Apparently, this example is in the draft of her NSF grant application. I learned of Johnson's NSF application for the first time when I got the email from you. I do not know whether that example made its way into Johnson's application before or after we wrote our Lebowitz Prize application. Johnson seems to think this example is hers and she is free to have it.[5] The example does not illustrate my notion of fine-tuning bias particularly well and it is only one of several examples I use. As I have already told Johnson, I will delete all discussion of PredPol in my paper.

6. I will now address also sentences highlighted in my paper and Johnson's work that did not make it into the "Text Comparison" document.

(a) In the highlighted version of my paper, there are three references to Johnson's NSF draft application which are not in the "Text Comparison" document.

In response: I do not know how sentences from the written version of my *Pacific APA* talk (which Johnson heard) made it into the draft of Johnson's NSF grant application, which I saw for the first time upon receipt of your email. Nonetheless, to avoid conflict and since I am not wedded to the specific wording in presenting basic facts about machine learning, I changed even these sentences.

---

[4] 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).

[5] It is worth noting that the Predpol example is discussed frequently. Here is an example: https://www.technologyreview .com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

(b) Johnson highlights my sentence "The algorithm itself is not represented but rather encoded in the system. There are reasons to think that the visual system operates in similar fashion on the basis of non-representational transformation principles" (p. 6) and this sentence from her dissertation: "In visual psychology, for example, theorists often describe the operation of various trans- formations within the visual system as abiding by rules or principles that are not ascribed to or represented in the individual." (p. 36f).

In response: I have not read Johnson's dissertation and I struggle to see how these two sentences are similar.

(c) I write "if there are problematic patterns encoded in the training data, then the predictions of the model generated by this training data will have those same problematic patterns" (p. 14). Johnson writes "If the data going into the training period are problematic, then we can expect the generalizations the program makes based on those data to be problematic as well" (Algorithmic Bias, p. 8).

In response: These two sentences express a basic fact about training sample bias. Similar such statements about patterns in training data generating problems in the ensuing algorithm occur in almost every article on trainings sample bias. Moreover, these two sentences present the relevant issue quite differently. My focus is on encoding and predictions, Johnson's is on generalizations. Nonetheless, I changed even this.

This difference mirrors the different contexts in which we are doing our research. Johnson's home is social cognition and the literature on implicit bias. She approaches issues about biased algorithms from that perspective. My home is perception and my work on biased algorithms is driven by research at the intersection between AI and neuroscience funded by a Mellon New Directions Fellowship (2019), a Guggenheim (2020), and a NEH fellowship (2021).

This concludes the discussion of alleged overlap with Johnson's work. As I hope to have shown, there is absolutely no basis for any of the accusations. Nonetheless, in the spirit of collegiality and in the hopes that this will accelerate finding a positive path forward, I have made changes in my paper in response to every accusation (except 6 (b), where I struggle to see any textual overlap).

***

In what remains, I discuss the accusations about textual overlap with three online sources.

7. "cross-race bias, which is the tendency to more easily recognize faces of the ethnic group that one is most familiar with…" (p. 4)

In response: In five sentences, there is significant overlap between my formulation and text from a *Wikipedia* article on cross-race effect. I appreciate this being brought to my attention. This was an honest mistake, for which I am deeply embarrassed. My only explanation for how this could have happened is that my procedure for using my notes was impacted when putting the finishing touches on this paper.[6] [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. This is not an excuse and I sincerely apologies for the oversight. I have rewritten the relevant section from scratch. I should note that these five sentences detail widely known examples of cross-race bias. They do not express original ideas.

8. "All models have some predictive error, given the statistical noise in the data, the incompleteness of the data sample, and the limitations of each different model type" (p. 8). And "Sample bias occurs when the distribution of one's training data doesn't reflect the actual environment that the machine learning model will be running in" (p. 13).

In response: These 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. There is significant overlap between them and sentences in two online sources. I am grateful for Johnson's detective abilities that alerted

[6] 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 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 read about them.

me to this mistake. These two sentences 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 the two online sources, but I must have read them at the time. I am deeply sorry this happened. I have corrected the mistake in the revised version.

4. Further Context

I conclude by adding background information that bears direct relevance to these allegations being raised in the first place. When working on our joint application in January 2021, Johnson expressed concerns about people stealing from her work at a frequency and with an intensity that goes well beyond what is typical even for a junior scholar. In each case, her allegations were unwarranted. She has sent emails similar to the one she sent me on September 9, 2021 to others in the profession expressing concern that their ideas were taken from her and asking to be cited. (I do not know whether she has made similar complaints to the Editors in Chief of other journals.) When we were having these conversations, it did not occur to me—clearly naïvely—that I would be the target of the kind of accusations she was launching against others. This preoccupation with intellectual ownership may be overriding the actual facts at stake here.

In our conversations, she was concerned not only about issues regarding intellectual property but was deeply aggravated even about the fact that several philosophers have recently started working on algorithmic bias. As I explained to her, not only is it important for our world if the issue of biased algorithms gains attention from philosophers; it is good for *her*: it will provide more opportunities for her to discuss her ideas and for her ideas to be discussed by others.

Thank you in advance for your time in considering this response. Please let me know if you have any questions and if you would like me to send you the revised version of my paper.

Yours Sincerely,

Susanna Schellenberg