Re: Concerns about research acknowledgement

Susanna Schellenberg <susanna.schellenberg@icloud.com>
Sat 9/4/2021 1:32 PM
To: Johnson, Gabbrielle <Gabbrielle.Johnson@ClaremontMcKenna.edu>

Hi Gabby,

Thanks for writing your email. I'm sorry but also saddened that you felt the need to do so.

Just a few thoughts. I started working on these issues long before I knew about you or your work. In particular the idea of bottom-up biases (which I don't take to be new to my or indeed your work) and the idea of similarities between such biases in natural recognitional and AI systems is something I've worked on long before I knew about you or your work. I talked about these topics with Gabriel Greenberg several times and he mentioned to me that he has a graduate student working on based algorithms. But I only got a chance to read your work shortly before we met to talk about your paper. I realized when we were meeting that you were this graduate student I had been hearing about. While I've been thinking about bottom-up biases and similarities between such biases in perceptual and AI systems long before I knew about you or your work, you have published on both topics before I have and I'll make sure to recognize your work accordingly. A bit more on bottom-up biases:

While I think the idea of bottom-up biases is overlooked in philosophical discussion, the idea is lurking around implicitly in all kinds of work in cognitive science. I don't think of it as something I came up with, but rather a real world phenomenon to which I'm drawing attention more explicitly than others have.

Given that according to you purely implicit biases are always influenced by what I'm calling top-down biases, I am not sure that what we're talking about is the same. But if you say your idea is the same as mine, I take your word for it and will cite you accordingly. In the context of our Lebowitz application we tried hard to present our ideas as having substantive overlap while also highlighting the differences between our ideas. As you might remember, I only realized when we were writing the application that there were similarities between your notion of purely implicit biases and my notion of bottom-up biases. Understanding it only then is entirely my fault and due to having read your papers less carefully than I should have. Either way, I'll make sure to recognize your work on this more explicitly in this paper and others.

The idea of fine-tuning biases came to me in March (long after we wrote the Lebowiitz application) and I recognized in March that the PredPol case is an example of fine-tuning biases. I will either use a different example to illustrate fine-tuning biases or if I continue to use that example, I will mention that I have it from you. To explain why I used the example in the paper without citation: In the paper, I follow what I take to be a widely-used practice of not citing authors on real life examples, but I can see why one should cite authors for such examples. To explain the similarity: I used the paragraph I wrote for the application on the basis of what you had sent me. I didn't go back to check how similar it was to what you had sent me.

As you know, of course, it is standard knowledge that there are three basic types of machine learning programs. I haven't compared the paragraph I wrote on this with ones in your papers. I will do so to make sure that the way I describe them is not too similar to the way you describe them.

I will continue to make sure to cite you wherever I can.

All best,
Susanna

On Sep 4, 2021, at 3:08 PM, Johnson, Gabbrielle <Gabbrielle.Johnson@ClaremontMcKenna.edu> wrote:

Hi, Susanna:

I'm emailing because last week I was asked to referee your "Varieties of Biased Algorithms" for*Philosophical Studies*. Based on the content of the paper, I knew that you are the author. For this reason and our general history, I declined the invitation to referee the paper, in order to preserve the anonymity of the review process. However, after reading the document, I couldn't help but notice similarities to my work that lacked appropriate acknowledgment. I remembered in the conversation we had over Zoom in April that you told me that you hoped I would come to you if ever I thought there were important similarities in our projects that I thought were opportunities for discussing my work or for clarifying how our views differ. So, I'm reaching out to you now to provide some specifics with respect to your paper.

Firstly, there is one paragraph about PredPol that is noticeably similar in exposition to a paragraph written by me for a project proposal I put together in November with a colleague of mine at CMC, Drew Schroeder, and that I shared with you (with his permission) in December for our joint document that we circulated to our Lebowitz Prize letter writers. Here's your paragraph:

> "More fine-tuning is not always better. 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. It is used by law 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 areas. On the basis of this historical data, the algorithm makes predictions about where police should be dispatched in anticipation of new crimes. If historical policing practices have been shaped by discrimination, then we can expect bias. 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."

And here's ours:

> "Machine learning programs trained on such data will therefore perpetuate those unjust patterns. An example of these problematic patterns can be seen in predictive policing algorithms like PredPol, which are used by law enforcement agencies across the country to identify potential "hotspots" for crime. Such algorithms rely on data in the form of historical records of the frequency of criminal activity in particular areas to make predictions about

> where police should be dispatched in anticipation of new crimes. Consider, then, what is apt to happen if historical policing practices have been shaped by discrimination. For example, if police have, due to historical patterns of racism, tended to over-patrol predominantly black and minority neighborhoods, then we can expect predictive software to continue to identify those neighborhoods as potential hotspots. It then dispatches police disproportionately to those areas, creating and collecting more data with which to continue the vicious cycle."

I am concerned by this degree of overlap, and so wanted to outline the issue to you now in hopes that you appropriately credit me before publication.

There are also a few other sections of your paper that are structurally quite similar, and, in some cases, again involving a number of explicit overlaps in exposition. Most notably, you have a section that describes the operation of a machine learning program (starting with "There are three basic types of machine learning programs…") that overlaps with my description from the draft of my Algorithmic Bias paper that I sent to you in December (the section starting with "Machine learning programs come in three basic forms…") And another notable similarity is in a passage of yours around footnote 9, where you give a description of proxy attributes that is nearly verbatim the description of proxy attributes that I give in my Algorithmic Bias and Structure of Bias drafts. You do cite me in the last of these, but it seems to me that these descriptions are based on my work and rely substantially on my own language. Therefore, they should be more fully acknowledged.

There are other, more substantive elements of the paper that concern me. For example, parts of the main theses seem to me to need more acknowledgment either by mentioning some of my previous work on such theses or in clarifying how your approach differs from mine.

Based on how you describe the project in the introduction, the paper has three primary theses: (1) there exists an important distinction between what you call "top-down biases", which are biases that stem from high-level beliefs of individuals and programmers, and what you call "bottom-up biases", which are biases that are data-driven and stem from low-level processing of a computational system; (2) "Bottom-up biases", you argue, "occur not only in AI but equally in cognitive and perceptual systems" and can occur at a range of levels in the computational hierarchy; And (3) among bottom-up biases, there are three different types: "fine-tuning bias", "feature-linking bias", and "training-sample bias".

Versions of your theses (1) and (2) play a central role in my work. For example, your distinction in (1) is substantially similar to the distinction I draw out in my 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 (a point I illustrate, like you, using a general-purpose learning algorithm). Your second claim (2) is the main thesis I take up in my Algorithmic Bias paper: that some data-driven biases occur just as well in artificial and natural cognitive systems. In these cases, it seems appropriate to either acknowledge that these derive from my work or include some discussion clarifying how your claims differ from my already published claims. Regardless, I would think that you should explicitly cite my work here.

You also consistently emphasize that bottom-up biases are ignored in the literature, stating

that they are “rarely recognized and barely understood” and that “biases are typically discussed in the framework of top-down biases,” without ever citing me as an exception. I was surprised by this because it is an omission that I brought up to you explicitly in our last Zoom conversation as one that worries me when I hear you make it, because it warrants acknowledgement that I am an exception.

Of course, you do cite me in three footnotes: (1) to say that I give a functional characterization of cognitive bias (but not that I motivate it using the distinction between high-level and data-driven biases), (2) to say that I argue (and you disagree) that biases are best characterized at the computational level of analysis, and (3) to cite me for discussion on proxy attributes. I’m glad that you do cite me in these three notes. But, given the similarities of the main theses and my antecedent knowledge that my work has influenced the development of yours, I’d like to ask that you include more substantial recognition of my contributions to the development of your thought on this and future work on these topics. I’m confident you recognize the value of senior faculty supporting junior faculty, and I appreciate that we can have an open conversation about these issues.

Thanks,
Gabby