Feminist Philosophy Quarterly

Volume 8 | Issue 3/4 Article 5

2022

Algorithmic Microaggressions

Emma McClure
Saint Mary's University (Halifax)
Emma.McClure@smu.ca

Benjamin Wald
Chiefs of Ontario
benjaminwald05@gmail.com

Algorithmic Microaggressions[1]

Emma McClure and Benjamin Wald

Abstract

We argue that machine learning algorithms can inflict microaggressions on members of marginalized groups and that recognizing these harms as instances of microaggressions is key to effectively addressing the problem. The concept of microaggression is also illuminated by being studied in algorithmic contexts. We contribute to the microaggression literature by expanding the category of environmental microaggressions and highlighting the unique issues of moral responsibility that arise when we focus on this category. We theorize two kinds of algorithmic microaggression, stereotyping and erasure microaggressions, and argue that corporations are responsible for the microaggressions their algorithms create. As a case study, we look at the problems faced by Google's autocomplete prediction and at the insufficiency of their solutions. The case study of autocomplete demonstrates our core claim that microaggressions constitute a distinct form of algorithmic bias and that identifying them as such allows us to avoid seeming solutions that recreate the same kinds of harms. Google has a responsibility to make information freely available, without exposing users to degradation. To fulfill its duties to marginalized groups, Google must abandon the fiction of neutral prediction and instead embrace the liberatory power of suggestion.

Keywords: algorithms, microaggressions, AI ethics, artificial intelligence, machine learning, bias

The last decade has marked huge advances in machine learning and the ever-increasing centrality of algorithms to our daily lives. Drawing on vast training datasets,

[1] We would like to thank McClure's colleagues at the University of Toronto's Mental Health and Disability Caucus for early brainstorming on the project, and Wald's colleague Gillian Hadfield, as well as the entire research community at the Schwartz Reisman Institute for Technology and Society, for their support and inspiration on this project. We are also deeply grateful to the organizers and attendees of the Feminism, Social Justice, and AI Workshop, and to the anonymous reviewers at *FPQ*, for comments that significantly improved the paper.

these algorithms can produce uncannily accurate predictions on a range of subjects, yet it has also become increasingly apparent that these systems can inherit the human biases present in this training data. Some systems have a large and obvious discriminatory impact, such as the COMPAS algorithm used to decide whether to grant people parole (Angwin et al. 2016), and other algorithms that determine who gets a loan or who is diagnosed with cancer (Osareh and Shadgar 2010). Other harms, however, arise from the accumulation of smaller and individually less impactful instances of bias, which nonetheless as a whole confront marginalized groups with prejudicial stereotypes and attitudes. The concept of microaggressions was developed to address exactly this sort of harm, and we will argue that the concept is still applicable and illuminating when the perpetrator of the microaggressions is software rather than human. Furthermore, we will demonstrate that recognizing these harms as instances of microaggressions helps us to avoid solutions that simply recreate the same kinds of harms, and adopting the microaggression paradigm will guide individuals, tech corporations, and governments towards a more just online world.

Section 1 will identify the kinds of cases where machine learning gives rise to microaggressions and situate these cases within the category of environmental microaggressions. In offline spaces, environmental microaggressions occur in “chilly climates” when features of the physical environment communicate hostility to and exclusion of members of marginalized groups (Sue 2010). We’ll argue that virtual environments can also be chilly climates: algorithmic microaggressions communicate that online spaces were not built for members of marginalized groups, that they do not belong in these spaces. We’ll further theorize two types of environmental microaggressions that are especially apparent in algorithmic contexts: stereotyping and erasure microaggressions.[2] Stereotyping algorithmic microaggressions present users with negative stereotypes about marginalized social groups, and they arise because the algorithm was trained on a dataset containing those stereotypical associations. Erasure algorithmic microaggressions, in contrast, arise because marginalized groups are underrepresented in the training dataset, and therefore, the algorithms trained on these datasets do not function, or function significantly worse, for marginalized groups.

Section 2 will take up the question of moral responsibility. As with other types of environmental microaggressions, numerous agents contribute to the creation of the microaggression—including those who create and train the algorithm, as well as those who contribute to creating the dataset the algorithm is trained upon, and those who use an algorithm that continues to learn after its original training—but we’ll

[2] During the review process, it came to our attention that similar discussions of environmental and erasure microaggressions can be found in Friedlaender (2021).

argue that tech companies who create the algorithms bear the most responsibility for the microaggressions they commit. Tech corporations often attempt to avoid moral responsibility for algorithmic microaggressions by claiming that their algorithms neutrally reflect biases that already exist in society—biases that the tech corporations are not responsible for. We'll give reasons to doubt this claimed neutrality. Tech companies are not merely mirroring preexisting biases; they're perpetuating oppression and perpetrating microaggressions.

Sections 3 and 4 will explore a case study: the stereotype microaggressions committed by Google's autocomplete search suggestions and the subsequent erasure microaggressions created by the recent policy changes Google made in response to criticisms of their autocomplete feature. The case study of Google's autocomplete demonstrates our core claim that microaggressions constitute a distinct form of algorithmic bias and that identifying them as such is key to effectively addressing the problem. We ultimately argue that to fulfill its duties to marginalized groups, Google must abandon the fiction of neutral prediction and instead embrace the liberatory power of suggestion.

1. Algorithmic Microaggressions

In this section, we'll show how algorithmic examples fit within the microaggression paradigm. Microaggressions are small slights that communicate demeaning and exclusionary messages about members of marginalized groups, thereby perpetrating and perpetuating their oppression.[3] While most philosophers

[3] In defining microaggressions in this way, we adopt a structural account of microaggressions, focusing on the functional role they play within an oppressive society (for similarly structural accounts, see Pérez Huber and Solórzano 2015, McTernan 2018, Friedlaender 2018, Freeman and Stewart 2018, McClure 2020). Thus we are departing from experiential accounts, like those held by Sue (2010), Fatima (2017), or Rini (2021), which would include as microaggressions only those examples where a member of a marginalized group is present and feels targeted by the (potentially) biased act. We also depart from psychological accounts, like Pierce (1970) or Friedlaender and Ivy (2020), which focus on the mental states of the microaggression's perpetrator. Although the algorithmic examples we'll discuss could potentially fit within either of these two other accounts, an experiential account wouldn't allow us to include examples like problematic autocomplete results shown to someone who is not a member of a marginalized groups (which we discuss in section 1.2); whereas a psychological account would have to argue that algorithms could have mental states such as conscious or unconscious bias. For further discussion of the differences between (and relative benefits of) structural, experiential, and psychological accounts, see McClure and Rini (2020).

and psychologists have focused on interpersonal microaggressions, which are communicated verbally or behaviorally, we'll focus here on environmental microaggressions, which are communicated by features of the environment (Sue 2010). Environmental microaggressions have traditionally been connected to physical features of the environment: buildings without wheelchair access, airplane seats that cannot accommodate larger bodies, or a wall displaying photos of past leadership that demonstrates only white people have been promoted and celebrated within a company. We'll expand the category of environmental microaggressions to include nonphysical spaces: in the virtual world, algorithmic microaggressions communicate similarly demeaning and exclusionary messages about members of marginalized groups. In section 1.1, we will give examples of algorithmic microaggressions and demonstrate how these algorithmic examples fall into two types: stereotyping and erasure microaggressions. Then in section 1.2, we will situate these algorithmic examples within the broader microaggression literature, highlighting the similarities between algorithmic microaggressions and other environmental microaggressions and showing how physical environmental microaggressions can be similarly separated into stereotyping and erasure subcategories.

1.1 Two Types of Algorithmic Microaggression: Stereotype and Erasure

First, let's start by surveying examples of algorithmic microaggressions. All these examples share the same basic structure as microaggressions: a pattern of results that would be individually innocuous collectively communicate hostility and exclusion. In the algorithmic context, these hidden hostile messages can often be made salient by comparing the disparate treatment of marginalized and privileged groups. A viral tweet in 2016 revealed that a Google image search for "three Black teens" returned a series of mugshots, whereas a search for "three white teens" returned images of smiling teens playing sports (Allen 2016). Similarly, Google autocomplete has a history of offering racist or sexist completions to innocuous queries. Carole Cadwalladr (2016) wrote about how starting off a query with "are Jews" returned as a prominent option "are Jews evil?", and if this option is selected nine out of the top ten results claim to show that yes, they are. Google changed their algorithm to prevent this result, but in 2018, an article in *Wired* showed that many problematic autocompletes remain, including "Hitler is my hero," "Blacks are not oppressed," and "feminists are sexist" (Lapowsky 2018). And the problem is not limited to Google. Facebook has an algorithm to determine which ads get shown to which users. A recent paper (Ali et al. 2019) has shown that the demographics that see an ad can change dramatically based on the content of the ad, often in line with stereotypes about those demographics. For instance, job ads for a lumberjack position were shown to an audience that was 90 percent male and 72 percent white, while a cashier position at a supermarket was shown to an audience that was 72 percent women, and a taxi driver position was skewed 75 percent toward Black viewers. More broadly, Caliskan, Bryson, and Narayanan (2017) have shown that machine learning based on textual data inherits the biases present in the text. In particular, machine learning systems showed the same kinds of associations as those found by the implicit associations test (IAT), with stereotypically Black names being more easily associated with unpleasant rather than pleasant words and the opposite for stereotypically white names, and stereotypically female names being more associated with family terms than career terms. Given the huge range of machine learning systems that are trained on this kind of data, we would expect (and indeed see) that these kinds of biases are widespread.[4]

In all these cases, the problem arises from the content that individuals are shown when they use various programs. More particularly, it arises from the algorithm's decision about what unsolicited content to show to people, based on predictions about what will be relevant. We will argue that these algorithms end up inflicting microaggressions when the prediction confronts the user with negative stereotypes about a marginalized group, as in the "are Jews evil" autocomplete result, or when they express negative stereotypes by presenting content based on harmful assumptions about a group the person belongs to—for example, if Facebook's ads assumed women were uninterested in jobs in STEM fields and only provided job ads to stereotypically feminine careers like secretary. Let's together call this form of microaggressions *stereotyping microaggressions*.

A second class of algorithmic microaggressions comes from technologies that simply do not function, or function significantly worse, for marginalized groups. An excellent example of this is the revelation by Buolamwini and Gebru (2018) that facial recognition technology is significantly worse at recognizing Black women's faces, with an error rate as high as 34.7 percent compared to 0.8 percent for white men. We can add to this the experience Black women have had of having their faces covered over by their Zoom backgrounds, or finding that automatic image previews on Twitter invariably centered white faces over Black faces in its cropped previews (Hern 2020).[5]

[4] To underscore both the importance and danger of this form of bias, Timnit Gebru was recently fired from Google's ethical AI team over an as yet unpublished paper on the ethical issues with large language learning modules. While the stated justification was that the paper was insufficiently sourced, this move is widely seen as an effort to suppress criticism of such modules (Metz and Wakabayashi, 2020).

[5] A related example is the failure of pulse oximeters to give accurate readings for people with dark skin, but this is not a microaggression of the kind we're interested in here—it does not rely on accumulated incidents to be harmful but is directly a matter of deep medical concern.

In these cases, the harm does not lie in being confronted with negative stereotypes. Instead, it lies in erecting barriers to members of marginalized groups for participating in certain spaces. While each individual instance may be a small annoyance, in aggregate these instances convey a message that members of marginalized groups are not welcome in spaces that rely on these technologies. Let us call this form of microaggression *erasure microaggressions*.

Both of these forms of microaggression arise from the way that machine learning systems use a set of training data. Machine learning refers to computer algorithms that are able to learn how to do a task without the need to explicitly program the necessary steps to achieve that task. These programs start with a set of data that the program will learn from, called the training set. Training a modern machine learning system requires a huge training set—the bigger the better. The most advanced models require truly vast training sets; the largest version of Google's state of the art language model GPT-3 was trained using a dataset containing 499 billion tokens. One component of the training data was all of Wikipedia, which made up 3 percent of the total data (Li 2020). GPT-3 is an outlier, orders of magnitude bigger than earlier language models, but even smaller language recognition systems require billions of tokens in the training set, and they are only getting larger.

Machine learning systems can learn patterns in the training set and use these patterns to generalize to new cases. Unsupervised algorithms find these patterns without any human pointing them out, and indeed, they can find patterns that humans are unable to recognize (Mahesh 2020). Even inspecting the algorithm and its outputs often still leaves us in the dark about what exactly they are picking up on. As an example, Google's AlphaGo, which was the first program to defeat professional Go players, didn't beat its human opponents by playing similar strategies and executing them better. Instead, it played in ways that differed dramatically from human play, picking up on patterns in its vast training set[6] that humans, even professional players of the game who have spent their lives looking for such patterns, didn't see.[7] Even with access to the inner workings of the algorithm, we cannot easily reverse engineer the pattern that is being tracked.

This inscrutability can cause problems. For example, Ribeiro, Singh, and Guestrin (2016) trained a neural network to distinguish between huskies and wolves.

[6] The original version of AlphaGo learned from records of human games, but the latest version used its own games as its training set—essentially playing itself millions of times and iterating its strategy based on this new data.

[7] Fan Hui, three-time European Go champion, described a pivotal move in a game between AlphaGo and Lee Sedol this way: "It's not a human move. I've never seen a human play this move"; and indeed AlphaGo itself calculated the odds of a human making the move as 1 in ten thousand (Metz 2016).

They intentionally selected photos of wolves with snow in the background and pictures of huskies without snow for the training set. Sure enough, the classifier learned to get the right answer on the training data by looking for snow, rather than looking at any feature of the animals themselves. With the huge datasets used for modern machine learning the failures are unlikely to be this obvious, but the risk is still present that they will fix on patterns that can fail to generalize to the real world. More real-world examples of such failures come from what are called "adversarial examples" (Goodfellow, Shlens, and Szegedy 2014). These are examples where by making tiny alterations that humans are unlikely to even notice, as small as a few pixels on an image, an otherwise successful machine learning system can be fooled into misidentifying an image. This becomes worrying if what it is misidentifying is, for example, a harmless pile of clothes as a gun in an airport scan, or a stop sign as a speed limit sign for a self-driving car. Just as with the much simpler example of snow versus no snow, the machine learning system has learned a pattern that can fail to generalize to new situations or, as we will see, a pattern that encodes biases and discriminatory stereotypes.

Another form of machine learning that can give rise to bias is the generative adversarial network, or GAN (Goodfellow et al. 2014). These systems consist of two machine learning models. The first model, the generator, tries to provide outputs that are as close as possible to the training data, while the second model, the discriminator, tries to detect which inputs are genuine and which were created by the generator. The two systems learn from each other over time, with the generator becoming ever more skilled at creating outputs that will fool the discriminator and the discriminator becoming ever more skilled at identifying these synthetic outputs. The point of such a system is to generate synthetic outputs that are as close to indistinguishable from real data as possible. These systems have been used to generate faces, detect anomalies in medical imaging, reconstruct images, and perhaps most worryingly, create so-called "deep fake" videos. GANs, like other forms of machine learning, can give rise to biased outputs. As an example, Jain and colleagues 2018 showed that a GAN tasked with coming up with synthetic images of engineers generated outputs that reinforced sexist and racist stereotypes (Jain et al. 2018).[8]

Machine learning training sets can give rise to microaggressions in two ways, corresponding to the two kinds of microaggression identified above. One potential issue is if the training data itself contains biases against marginalized groups. GPT-3 was trained largely using data gathered from the internet and from digitized books. These models work by associations—more complex versions of the autocomplete

[8] See Tschaepe (2021) for further discussion on biases that can arise from GANs and how we can understand them through a pragmatist lens.

feature on emails that guess the most likely words to follow the words you have already typed. So if the training data contains harmful and stereotyped associations towards marginalized groups, these associations will be inherited by the machine learning model. This is the source of stereotyping microaggressions, where the outputs of the machine learning model replicate and confront people with the biases already present in society.

On the other hand, sometimes the problem is not the presence of particular biased associations within the training data but rather the absence of marginalized groups from the dataset. For example, if the dataset on which facial recognition software was trained contained only a small number of Black faces, the system will be worse at recognizing such faces. So erasure microaggressions arise from the absence of representation within the dataset.

All the particular examples discussed have been recognized and criticized previously as problematic instances of bias. However, they have not previously been seen as instances of microaggression. In the next subsection we will show why it is both accurate and enlightening to understand these instances of machine learning bias through this lens.

1.2. Situating Algorithmic Microaggressions within the Broader Microaggression Literature

We've already highlighted some clear similarities between the above examples of algorithmic bias and microaggressions. Most obviously, the similar pattern of aggregation: one microaggression, cringey autocomplete result, or imperfect Zoom filter is easy to brush off, but a lifetime of repeated stereotypes and erasures can accumulate into serious harm to members of marginalized groups. We'll now delve deeper into the similarities, which will allow us to further specify where algorithmic microaggressions belong in the microaggression taxonomy, and we'll also demonstrate the advantages of expanding the microaggression concept to include these algorithmic examples.

We proposed that algorithmic microaggressions are a type of environmental microaggression. Rini (2021, 21) defines environmental microaggressions as "background facts that regularly confront marginalized people with casual disregard or disdain," and Sue (2010, 25) points out that such microaggressions are often discussed by other names: "When people refer to the 'campus climate' as hostile and invalidating, or when workers of color refer to a threatening work environment, they are probably alluding to the existence of environmental microaggressions." Crucially for our purposes, environmental microaggressions aren't perpetrated by individual agents, or even by cohesive groups of people with a common goal. Instead, the microaggressive hidden message is communicated by the physical space itself. To draw a few examples from the philosophy literature:

- A wall of photos celebrating a faculty that includes few or no people of color sends the message "You don't belong here" to BIPOC faculty and students (Henning 2020, 258).
- Airplane seats that cannot accommodate a fat person send a similarly exclusionary message, compounded by the fact that fatphobia is reinscribed everywhere from X-ray machines to the fashion industry (Reiheld 2020, 208 and 213).
- When all the bathrooms in a venue are gender-specific or when a medical intake form only includes checkboxes for male/female, trans and nonbinary people are told "There is no box for me" (Dean, Victor, and Guidry-Grimes 2016, 561).

The above environmental microaggressions are parallel in structure to the erasure algorithmic microaggressions that we introduced in the previous subsection. No single person created a room full of photographs featuring exclusively white faces or chose to design airplane seats with the intention of making fat people uncomfortable—just as no one person encoded discriminatory content into Google's autocomplete algorithm. Instead, the exclusionary message is sent by recurring features of the environment—ranging from architecture to interior decorating to bureaucratic paperwork. When we turn to nonphysical environments, algorithms could be said to function as the architecture and bureaucracy of the online world, creating unwelcoming and hostile spaces that send the message "Members of marginalized groups don't belong here."

Combining the algorithmic and physical examples of erasure microaggressions also brings into view a feature of environmental microaggressions that has not yet been much discussed: environmental microaggressions are still microaggressions even if no member of the targeted marginalized group is present to witness them. Here we expand from Reiheld (2020, 207), who discusses how antifat environmental microaggressions, like not being able to fit comfortably in an airplane seat or a medical imagine machine, also send harmful messages to those who are "not yet fat" by encouraging self-disciplinary tactics aimed at remaining thin. Reiheld (2020, 218) argues that microaggressions "cause all persons to fear being fat, including those who are not," and fatphobic microaggressions don't have to directly target fat people in order to contribute to their continued marginalization. We extend Reiheld's argument further to include the other environmental and algorithmic cases we have been discussing: these microaggressions simultaneously erase and normalize erasure, whether or not a member of the marginalized group is present to witness their exclusion. In physical space, a company can't excuse its wall of white photos by saying there are no employees of color to be offended by their lack of representation—the absence of people of color is both cause and consequence of the hostile workplace climate. Similarly, in online space, Twitter's image-cropping algorithm isn't only a microaggression when it is presented to members of the racialized groups it decenters; it is also a microaggression when it is presented to a white user, whom it centers. In both cases, the oppressive message is still sent—whites are the default focus; the absence of people of color is normal—and the cycle of oppressive absence is reinforced. Marginalized people are erased from our workplaces, our screens, our lives, and privileged people don't even notice they are missing.

Turning now to stereotyping microaggressions, we can also find offline examples that share the same structure:

- The (now discontinued) practice of naming hurricanes exclusively after women reinforced stereotypes that women are irrational and destructive forces of nature (Rini 2021, 21–22).
- US medical school curriculums that only mention the health needs of gay men when discussing the risk of sexual transmitted diseases feed into stereotypes of gay men as promiscuous and sick (Dean, Victor, and Guidry-Grimes 2016, 562).
- Sports mascots and logos, like the Redskins, depict Indigenous peoples as savage and primitive (Sue 2010; Steinfeldt, Hyman, and Steinfeldt 2019).

These environmental microaggressions clearly stereotype members of marginalized groups, yet the demeaning stereotype is not transmitted by any particular individual. Instead, the symbol or practice, created and sustained by numerous, loosely connected individuals, conveys the stereotyping microaggression, and furthermore, people may participate in producing the stereotype microaggression without being aware of their participation. For instance, the practice of depicting Indigenous peoples in derogatory ways originated long before the Redskins adopted their logo, and the creators of these earlier depictions unknowingly created the context for the current stereotype microaggressions (Corbett 2019). Similarly, for stereotyping algorithmic microaggressions, numerous coders and content creators contribute to producing the stereotyping outputs. For example, the original authors of the Enron emails had no idea that their emails would later be used in the machine learning dataset that trained Natural Language Processing (NLP) algorithms (Levendowski 2018), yet they did contribute. As we'll discuss in more detail in the next section, the existence of these unknowing contributors makes assigning moral responsibility for algorithmic microaggressions a particularly fraught project.

Before we turn to the question of moral responsibility, however, we would be remiss not to mention the work of Chester Pierce, the first microaggression theorist. Although he does not use the term environmental microaggression (which was coined by Derald Wing Sue after Pierce's retirement), Pierce discusses the pattern of stereotypical portrayals of Black people in film and television in *Psychiatric Problems of the Black Minority*:

> The mass media more often than not see to it that blacks are portrayed in ways that continue to teach white superiority. . . . For instance, a black is more often the server than the served, for example, on a commercial the black pumps the gas while the white drives the car or the black woman is the cab driver while the white man's uncivil remarks give her a headache. The black can be predicted to be less often depicted as a thinking being . . . the black is seen over and over in such guises as a server and a non-thinking physical creature. (Pierce [1974] 2015, 10–11)

Pierce's examples are particularly illuminating because together they demonstrate the compounding effects of stereotyping microaggressions. Any individual commercial, TV show, or film that portrays Black Americans as servile and unthinking is individually problematic (and also the creation of numerous individuals from pre- to post-production), but when we view all these problematic portrayals together, an even more powerfully damaging pattern emerges. After the quoted passage, Pierce goes on to explain how consuming the constant barrage of stereotyping microaggressions could brainwash Black youth into unquestioning acceptance of their limited, unsatisfying roles in society. Moreover, in Pierce's (1970) earlier work, "Offensive Mechanisms", he demonstrates that privileged people are also affected by pervasive stereotyping microaggressions. He argues that white children who witness anti-Black microaggressions learn to expect deference and presume their own superiority: "Society is unrelenting in teaching its white youth how to maximize the advantages of being on the offense toward blacks" (Pierce 1970, 269–70).[9]

Applying Pierce's insights to the algorithmic examples, we can see even more clearly how the harm of stereotyping microaggressions compounds. One Google autocomplete search—and all the individual content creators who contributed to the stereotyped autocompletion—is somewhat problematic, but when we consider all searches, on all platforms, and all the other stereotyped portrayals in the online world, we begin to understand just how damaging stereotyping algorithmic microaggressions can be. Moreover, we can see how the damage is done to both members of the targeted population and to the rest of society, who have their

[9] Of course, neither we nor Pierce are suggesting that this continual pressure is a form of oppression. Being taught to assume white superiority often materially benefits white people, even if it also harms their epistemic or moral capacities (Mills 2007).

stereotypes reinforced and are continually taught to recreate these offensive mechanisms.

2. Moral Responsibility

Now that we've shown how algorithmic examples fit within the microaggression paradigm, we're ready to consider the unique issues of moral responsibility raised by algorithmic microaggressions. For physical environmental microaggressions, there is often a clear institution that bears (or should take) responsibility. In the examples we surveyed in section 1.2—the workplace that put up the photo wall, the airline with the too-small seats, the health-care provider with the discriminatory intake form, and the football club with the racist logo—all seem clearly responsible for the microaggressions they've perpetrated, or at least for changing policies in order to avoid perpetrating similar microaggressions in the future.[10] But as we'll show in this section, the story of moral responsibility is not as simple for the machine learning processes that create algorithmic microaggressions. We'll focus on two methods that tech companies have used to avoid taking responsibility for their algorithms—moral proxy and neutrality—and give reasons to reject each.

The first complication is that tech companies can, with some accuracy, claim that the machine learning systems we are discussing reflect bias that is already present in society. At least in those cases where the training data is in fact representative, the biased results arise from the algorithm learning patterns that are actually present in the data and accurately reproducing them. So we might ask who is truly responsible for the harmful microaggressions inflicted by biased algorithms: is it the company that built them, or the biased individuals whose data created the algorithmic bias, or perhaps society in general? This is a version of what is called the moral proxy problem—since AI's are not themselves morally responsible agents, who is the appropriate moral agent to bear responsibility for morally relevant actions taken by the AI?[11]

One possibility is that this is an instance of a "responsibility gap." Andreas Matthias (2004) argues that one of the problems with what he calls learning automata, which would include the AI techniques used in many instances of algorithmic microaggressions, is that they give rise to responsibility gaps, in which no one is truly responsible for a harm.[12] This is because no one satisfies the relevant

[10] For more on institutional moral responsibility for nonalgorithmic microaggressions, see Brennan (2013, 2016) and Dean, Victor, and Guidry-Grimes (2016).

[11] See Millar (2015), Himmelreich (2018), Köhler (2020) and Thoma (2022) for further discussion of the moral proxy problem in AI

[12] See also Sparrow (2007) for an application of this idea to autonomous weapon systems, and Danaher (2016) for the related idea of a retribution gap.

control condition required for moral responsibility. Neither the user nor the designer has sufficient control over the relevant actions of the automated system to count as responsible for the outcome, and so neither cannot be held responsible.

However, this worry misconstrues the kind of control that is necessary to count as morally responsible. As Sebastian Köhler (2020) argues, we can understand the kind of responsibility we have towards AI on the model of the responsibility we have when we make use of other supervised minimal agents, such as nonhuman animals. Thus, he argues, the appropriate model for understanding this responsibility is that of using another agent as a tool. We clearly lack full control over the actions of nonhuman agents whom we utilize as tools, but just as clearly we are often responsible for their actions. In the same way as the decision to use a nonhuman animal, and the role humans have in training it, makes those who choose to deploy such agents morally responsible for their actions, people choose to deploy AIs and control the training that they are given; so even if the results are never perfectly predictable, programmers and corporations still bear the responsibility for the outcomes their AIs create.

We might instead argue that the user is the one truly responsible for the outcome. There are various arguments in, for example, the ethics of self-driving cars that hold that the user of the car is the one who should be held ultimately responsible for harm caused by the car (Nyholm 2018). Generally, this is considered in cases where the car causes injuries to others rather than the occupant themselves, as it is odd to hold someone morally responsible for harm to themselves. Still, we might hold them responsible in the same sense as we hold someone responsible for harm they cause themselves with a tool, where responsibility does not necessarily come with moral blame but does absolve others of moral responsibility for the outcome. In the same way, we could argue that the user of the AI technology that causes the microaggression is ultimately responsible for the outcome, and so absolve the producer of the technology of responsibility.

This line of argument is most compelling when the user has a degree of choice over how the system operates. For example, if individuals have the choice to adopt one of several "ethics settings" for a self-driving car, some of which are more altruistic and favor protecting others and some of which are more selfish, prioritizing the user, then it makes sense to consider the user as having responsibility for the outcome. However, there is no such choice available in using the kinds of AI systems that give rise to microaggressions.

We could still argue that there is a relevant choice: the choice to use the AI system in the first place. However, this is in many cases merely the illusion of choice. AI systems are becoming increasingly ubiquitous, and avoiding them becomes ever more difficult. Furthermore, many of these systems are incredibly valuable and time saving. To require minoritized populations to choose between experiencing microaggressions and forsaking the use of these technologies altogether would be an injustice to those who are already most vulnerable.

Finally, we might argue that the appropriate target of moral responsibility is in fact the individuals whose biased data has contributed to training the AI system.[13] This has been a popular approach, since it allows companies to avoid taking responsibility for moderation.

Big tech companies have been reluctant to explicitly adopt the role of content moderators for their algorithms and platforms. It is likely that Google, for instance, would be reluctant to intentionally sculpt the data fed into their algorithms in order to prevent microaggressive recommendations. One strong reason for this is that it would require Google to take firm stances on controversial topics surrounding issues like police accountability, trans rights, and other areas where any stance they take will inevitably invite criticism. Much better, from a public relations perspective, to have an algorithm that neutrally reflects the data of users without taking sides—this way they can deflect criticism from both sides of controversial issues while intervening on an ad hoc basis in areas where public opinion is sufficiently settled that there is less risk of blowback. And there is some public support for this claim of neutrality; while a slim majority of people in a Pew study agreed that algorithms will always reflect human biases, 40 percent still thought it was possible for algorithms to be neutral (Smith 2018).

While there is certainly enough responsibility to go around, we want to push back against the idea that the companies that design these problematic algorithms are neutral transmitters of the users', or of society's, views.[14] The neutral transmitter excuse assumes that it is possible to design a fully value-neutral algorithm, whereby the design of the algorithm makes no value judgments and prioritizes no particular viewpoint. But this is not in fact possible, as we will argue.

We would argue that there is no such thing as a purely neutral way of organizing or displaying information. Even simple graphs make numerous decisions about how to frame information that require value judgments. As Catherine D'Ignazio and Lauren Klein (2020) argue in *Data Feminism*, the idea of an objective framing of information is a form of the "view from nowhere" that pretends that we can possess

[13] This will only be an option for stereotyping microaggressions, not erasure microaggressions, since in erasure microaggressions it is not the presence of bias but the absence of diversity that explains the outcome. Still, stereotyping microaggressions are a significant enough range of microaggressions to be worth considering this argument.

[14] Safiya Umoja Noble (2018) similarly critiques the United Nations advertising campaign that raised sexist and racist autocomplete results for focusing the blame on users of search engines rather than the search engines themselves.

a disembodied perspective, a “god’s eye view,” untainted by our subjective experience. Machine learning algorithms are immensely complex, and this very complexity can obscure the value judgments being made in the design of the algorithm. Let’s take as an example Google’s page-rank algorithm, which determines the order in which results are displayed when you do a search. The idea is that the top results will be the ones that are most relevant to your query. But once we examine this notion, it is clear that we will need to operationalize the idea of relevance. One easy way to see that it is impossible to measure relevance itself, free of interpretation, is the huge industry of search engine optimization (SEO). SEO professionals work to help boost web pages up the ladder of Google search results, complete with a division between “white hat” optimizers, who follow search engine guidelines, and “black hat” optimizers, who try to exploit loopholes and otherwise break the search engine’s guidelines (Patil, Pawar, and Patil 2013). This industry thrives because the quality, or relevance, of a page is subjective, and there is a lot of money to be made by identifying and matching Google’s (or another search engine’s) own definition and proxies thereof. Furthermore, the influence of advertisers already shows that the ranking used serves a specific set of values, rather than neutrally reporting some underlying truth. As Safiya Umoja Noble (2018) points out, commercial values are clearly driving the position of the “sponsored” results, and most users of Google cannot reliably identify which results are advertisements and which are the result of the page-rank algorithm.

But stepping back a bit, we can see that this is unsurprising, since it is impossible even in principle to avoid value judgments in algorithms. As Gabbrielle Johnson (forthcoming) points out, an algorithm must balance competing considerations against one another. Perhaps we need to trade off risk of error against ease of use, as when an aggressive autocomplete algorithm automatically changes an unfamiliar word that you had in fact spelled correctly into another word entirely. Deciding how to make these trade-offs involves value judgments. But, as Johnson points out, not only are these themselves values, we also cannot cleanly demarcate these values into epistemic and moral values. Prioritizing ease of use over avoiding errors, for example, helps those who most cleanly fit into the expected boxes for the product, while those on the periphery will be exposed to the additional errors. Johnson draws a parallel between this and feminist criticisms of the value free ideal in science. Longino (1995), for example, argues that the choice between novelty and consistency as epistemic values in a scientific theory is itself laden with moral values since consistency favors the status quo, which in turn favors the current patriarchal system. And that is before we even consider the arguments that moral values are themselves epistemic values, as some defenders of pragmatic encroachment claim.[15]

If algorithmic design does indeed always include value judgment, then the choice is not between neutrality, on the one hand, or choosing to shape algorithms according to our moral judgments, on the other. Rather, it is a choice between two sets of moral judgments. As we will see in the next section, the current set of choices being made, and the options they provide for addressing microaggressions, are insufficient and need to be improved.

3. Google Autocomplete: A Case Study in Algorithmic Microaggressions

The current approach of many companies whose algorithms inflict microaggressions is ad hoc and reactive. Furthermore, since companies do not currently conceptualize the harms of algorithmic bias as specifically microaggressive, seeming solutions can do more harm than good. We will use Google autocomplete as our main example, since it has been widely studied and provides a clear example of the kind of approach we want to criticize. However, it is one thing to say that Google and others who use similar algorithms should do better; it's another to specify ways in which they could improve. In this section, we will point out several shortcomings with Google's current response to algorithmic microaggressions: (1) Google's response is ad hoc, correcting autocomplete outputs based on whether they cause offense, when it should instead be considering the functional role of microaggressions within structural oppression, (2) Google's attempted solution, removing autocomplete suggestions, risks creating unjust barriers to access of information for members of marginalized groups, and (3) Google's attempted solution also ends up creating a different microaggression—replacing a stereotyping microaggression with an erasure microaggression. In the next section, we will lay out a more promising, multifaceted approach.

To its credit, Google has made some attempt to respond to their algorithm's microaggressions—once they were called out. Noble's (2018) book, *Algorithms of Oppression*, brought the problem of autocomplete to the public's attention. The cover features the query "Why are Black women so . . ." with predictions such as "angry," "loud," and "lazy." Noble's book effected change, and quickly: by 2019, Google had completely revamped its autocomplete policies. Yahoo still to this day autocompletes "Why are Black people" with terms like "violent" and "inferior," and "Why are

[15] See Basu (2019). The idea is that moral and other pragmatic considerations are directly relevant to (encroach upon) purely epistemic questions such as what to believe. So, for instance, even if my evidence supports assuming that the one Black man at a golf club is an employee, moral considerations properly tell me to suspend judgment due to the moral harm in making an error in this context.

women" with "bitches" and "crazy,"[16] but Google has made significant strides by allowing users to report "inappropriate predictions" for being "hateful against groups" (Google 2021).

Many of the specific examples that have been widely publicized (by Noble and others) have since been corrected; often by drastically reducing the number of autocomplete suggestions offered. For example, at the time of writing, a search for "Why are Jews" returns only three rather than the usual ten autocomplete suggestions, presumably so Google can avoid having anti-Semitic results show up. Google's autocomplete policies (Google 2021) lay out classes of predictions that they do not provide, including sexually explicit, vulgar, hateful against groups, and sensitive and disparaging terms associated with named individuals. However, while they have made some progress addressing the individual search terms, or classes of search term, that people have called attention to, many problematic search results remain. A recent study by Roy and Ayalon (2020), for example, found that searches related to older women returned many more negative and disparaging autocomplete results than did the same searches about older men and that the autocomplete results in general displayed a negative impression of the elderly as a whole.

It is difficult to talk about the exact procedure used to remove or curate autocomplete results since Google is not transparent about these procedures. However, we know that autocomplete results that gain negative publicity are fixed but that new examples can readily be found. This seems to resemble the approach to eliminating nonalgorithmic microaggressions by providing lists of things not to say or topics to avoid. For example, DiversityInc (2016) provides a list of nine things not to say to female coworkers, including asking if they are pregnant, calling them emotional, and telling them to be tougher.[17] Similar lists are often used in sensitivity and diversity training (Dean, Victor, and Guidry-Grimes 2016). While presumably well-intended, these lists are not effective in eliminating microaggressions because they don't provide guidance for novel examples, especially less-frequently discussed forms of bias, since they don't take into account the root of the problem: structural oppression.

This brings us to the first problem with Google's solution. As we've discussed in section 1, algorithmic and other microaggressions are not wrongful because they cause offense; they are wrongful because they participate in and perpetuate structural oppression. Thus, unpopularity is not a reliable metric for tracking which phrases are microaggressions. As we've seen, removing unpopular results from autocomplete (or speech) won't remove all the microaggressions—microaggressions

[16] Author's search, July 2021.

[17] While the article doesn't use the term microaggression, the examples given are classic examples of microaggressions.

against Black women and Jewish people may be reduced while microaggressions against the elderly remain untouched. Moreover, removing all unpopular results ends up removing results that are not microaggressions. At the time of writing, "Why are white men" returns zero results,[18] even though white men are not a group of people at risk of structural oppression (though individual white men may belong to other social groups that are at risk). As long as Google is tracking unpopularity instead of microaggressions, they won't catch all (or only) the problematic autocomplete results.

Secondly, even if Google could eliminate all instances of stereotyping algorithmic microaggressions for all oppressed groups, they would just create new problems. Commonly, removing stereotypical autocomplete results is done by removing all the results for those search terms or severely limiting them. For example, returning to one of the earliest examples of this problem, typing "Why are Jews" into Google returns only three autocomplete suggestions ("kosher," "the chosen people," and "circumcised") rather than the ten that are offered for most results; whereas "Why are Black women" and "Why are lesbians" return no autocomplete suggestions at all.[19] While this does avoid exposing members of marginalized groups to negative stereotypes about themselves, it also inhibits their ability to access resources that would help them counter those stereotypes. Searches for information about their social group must be done without the aid of the time-saving algorithms such as autocomplete—a small cost to pay on any one search, but when that cost accumulates across multiple searches by many users, it adds up. Google's promotional material claims that autocomplete saves users "over 200 years of typing time per day" (Google 2021), and we argue that this time-saving benefit should be shared by Black women and Jewish people attempting to access material that would help them resist internalizing harmful stereotypes.

We further argue that inhibiting access to this information amounts to a form of epistemic injustice; specifically, what Miranda Fricker (2007) calls "hermeneutical injustice." [20] Hermeneutical injustice occurs when a member of a marginalized group is unable (or significantly inhibited) from understanding their experience of

[18] Author's search, July 2021.

[19] Author's search, July 2021. "Why are lesbians" further recommends turning on safe search—presumably to avoid seeing pornography in search results—but we would argue that the suggestion of safe search is an additional microaggression, when we consider how this association recalls the history of hypersexualization of lesbians, as well as current attempts to ban 2SLGBTQ+ content in schools because it is "too adult" for children to be exposed to.

[20] See Fatima (2017) for more connections between microaggressions and epistemic injustice.

marginalization because of gaps in the conceptual resources available to them. Fricker (2007) gives the example of women in the 1960s and ’70s who found it difficult to conceptualize their experiences of workplace sexual harassment, until they participated in consciousness-raising groups and came to the realization that many of them were experiencing the same mistreatment, which they then named as sexual harassment. Talking about their experiences allowed these women to counter the hermeneutical injustice and achieve self-understanding (as well as the ability to organize politically and resist further harassment). Nowadays, the internet has served as a powerful hermeneutical equalizer—allowing queer children and teens from small, conservative towns to learn “It gets better” (itgetsbetter.org) and enabling sexual violence survivors to continue the consciousness-raising projects Fricker discussed, through hashtags like #MeToo, #Time’sUp, and #MMIWG. Concepts that were previously unavailable or extremely difficult to access for members of these marginalized groups are now accessible to anyone with internet access. In fact, as Torino and colleagues point out, the very concept of microaggressions has been popularized and rendered accessible by the internet (Torino et al. 2019). However, internet users still have to know what terms to search for to find these resources, and here’s where we come back around to our worry about Google’s autocomplete: turning off the autocomplete for searches on “Why are Black women” or “Why are lesbians” inhibits members of these multiply marginalized groups from learning about concepts that are necessary for understanding their marginalization—or even their physical health risks. Of course, they are not completely prevented. They could still learn about misogynoir and homophobia, or their higher risks of dying in childbirth or contracting breast cancer, by typing in those search terms themselves, but first they would have to know to look. Autocomplete could help smooth the way towards self-understanding, but instead, with autocomplete turned off, it takes more time and more background knowledge to find this information. While any one instance of such an injustice will be minor, we suggest that this is a kind of epistemic micro-injustice, where the cumulative effect of many small inconveniences makes certain kinds of self-knowledge more difficult to acquire for marginalized groups.

Caution is in order here. As Gaile Pohlhaus Jr. (2012) points out, what looks like hermeneutical gaps in an oppressed group, from the point of view of the privileged, may instead be cases of willful hermeneutical ignorance on the part of the privileged. Willful hermeneutical ignorance occurs when individuals in an oppressed group are perfectly capable of articulating their oppression, but the privileged group fails to acknowledge and take up these hermeneutical resources and so renders the experience of the marginalized unintelligible to themselves. The privileged group uses faulty hermeneutical resources, and they are culpable for doing so because they adopt these resources out of prejudice.[21] Because of these faulty hermeneutical resources, members of the privileged group systematically misinterpret and fail to recognize the experiences that members of the oppressed group communicate to them. With Pohlhaus's critique in mind, identifying specific cases of hermeneutical injustice from the point of view of the privileged group is fraught with risk, since we may be identifying our own ignorance and inability to accept the hermeneutical resources that are present rather than a genuine gap in hermeneutical resources available to the oppressed group. Still, while identifying specific cases is best left up to members of the oppressed group, we can identify the possibility of such hermeneutical gaps and the possibility that eliminating autocomplete can contribute to or perpetuate these gaps.

Now we have set the stage to see the third problem with Google's autocomplete: when autocomplete works better for privileged groups, who don't need to overcome hermeneutical injustices, than for marginalized groups, who do face those barriers, then autocomplete becomes the site of an erasure microaggression. The lack of autocomplete suggestions on "Why are Black women" and "Why are lesbians" means members of marginalized groups see themselves less represented in online space, sending the microaggressive message, "This technology was not built for members of my social group."[22] In other words, in removing stereotyping microaggressions from autocomplete results, without replacing those results with suggestions of nonstereotypical and counterstereotypical searches, Google has turned a stereotyping microaggression into an erasure microaggression. Perhaps this is progress of a kind, but it is far from a satisfactory solution.

Before we suggest our preferred avenue for progress on these problems, let's consider one solution that we won't be recommending: Google could avoid creating erasure microaggressions by expanding the circle of search results that do not provide any autocomplete options—at the limit, by forgoing autocomplete suggestions altogether. Eschewing the use of machine learning algorithms may sometimes be the best way of addressing the problems they raise. A good example of this is Twitter's

[21] See also Dotson (2011) for a discussion of privileged groups' culpability for pernicious ignorance.

[22] As we've mentioned, "Why are white men" also yields no search results, but since white men are not marginalized (qua their whiteness/maleness) they are not structurally positioned to experience hermeneutical micro-injustice or erasure microaggressions. They might, however, experience epistemic harm—in not being able to access information about white ignorance (Mills 2007) that could help them to acquire self-knowledge about the ways they have participated in oppressive structures—but as Frye (1983) famously argued, harm is not the same as structural oppression.

image-cropping algorithm, which was designed to automatically crop photos uploaded to Twitter to display the most "salient" regions of the photo. This algorithm drew criticism for centering white faces more often than Black faces when both appeared in the same image.[23] After their own research confirmed the presence of a bias in the algorithm, Twitter discontinued the feature altogether. As they put it in on their own blog, "We considered the tradeoffs between the speed and consistency of automated cropping with the potential risks we saw in this research. One of our conclusions is that not everything on Twitter is a good candidate for an algorithm, and in this case, how to crop an image is a decision best made by people" (Chowdhury 2021).[24]

However, this approach is unlikely to be a satisfying response across the board. All machine learning algorithm systems are vulnerable to bias, so if this is our only way to address bias, then we will end up needing to abandon all algorithmic recommender systems. In some cases, it is hard to see how we could do without these systems; without search engines, the internet would be practically inaccessible, and all search engines rely on machine learning algorithms. In the case of autocomplete, we could do without it, but there would still be a significant cost: as mentioned before, Google claims that autocomplete saves users over two hundred years of typing time per day. Furthermore, these benefits are especially significant for those with disabilities that make typing difficult, so the decision to eschew these benefits would hit this group of people especially hard—creating a further barrier to equitable access and risk of erasure from online spaces.

Finally, to do without autocomplete would be to miss out on potential benefits of alternative solutions. As we will propose below, a better solution to the problem could not only avoid bias but actively fight it by providing counterstereotyping results as autocomplete options. We probably can't, and shouldn't, put the genie of machine learning back in the bottle, but Google still can, and must, do better.

4. A Better Way Forward for Google Autocomplete and Other Algorithmic Microaggressions

In the last section, we used the Google autocomplete case to demonstrate how conceptualizing some forms of algorithmic bias as microaggressive can allow us to better identify problems, including problems created by attempted fixes. In this final section, we'll show that the microaggression paradigm also guides us towards better solutions. We'll begin with solutions that have already been suggested by

[23] This is one instance within a long pattern of photographic technology performing poorly for Black users. See Benjamin (2019) for more on this discriminatory history.

[24] See also Yee, Tantipongpipat, and Mishra (2021) for more details on the research Twitter did to test for and identify the bias in their recommender system.

microaggression theorists to combat environmental microaggressions. We'll show how these preexisting solutions could be adapted to the algorithmic context, and then we'll conclude by suggesting two possible solutions unique to algorithmic microaggressions. Individually, these potential solutions will be limited in their effectiveness, but combining these different approaches—some individuals, corporations, and governments working in tandem—would create more genuine progress on reducing microaggressions in online spaces.

The first solution we'll explore is inspired by the work of Chester Pierce. As we mentioned in section 2, Pierce devoted much of his research to studying the effects of environmental microaggressions (under a different name) in television, film, and commercials. In fact, he didn't just research these effects—he took steps to combat them by getting involved in the entertainment industry. He consulted on a new children's show, *Sesame Street*, that was created in order to close the education gap between white and racialized pre-K students (Greene 2019). *Sesame Street* is most well known for teaching reading and math skills, but under Pierce's guidance, it also imparted a "hidden curriculum" of racial equality and self-respect for members of marginalized groups (Harrington 2019). The show featured a racially diverse cast that included Black authority figures, Gordon and Susan, and Black children who were smart and capable. These characters served as counterstereotypical exemplars that insulated Black children in the audience (and their parents) from the microaggressive messages sent by other media.

Pierce chose to research media microaggressions and consult on *Sesame Street* because of the ubiquity of televisions. If Pierce were alive today, he might take a similar approach to combating microaggressions in the now pervasive online context. Rather than attempting to control Google's algorithm directly, he might focus instead on creating content that would support racial equality and self-respect for members of marginalized groups. The impact of algorithmic microaggressions would be lessened if search results returned websites filled with counterstereotypical exemplars. For instance, a Google search for "Black men" currently returns as its top hit an article entitled "Outstanding Black Men in Canada 2020."[25] The article appears in *Shifter* magazine, which describes itself as "a Canadian online Black and urban culture magazine celebrating the best in music, film, television, fashion and sports."[26] Pierce would have recognized that websites devoted to representing Black equality and excellence can impact online racism and reduce the dangers of internalizing algorithmic microaggressions. Although no single individual creates algorithmic

[25] Kevin Bourne, "Outstanding Black Men in Canada 2020," *Shifter*, June 21, 2020, https://shiftermagazine.com/shifters/outstanding-black-men-in-canada-2020.

[26] "About," *Shifter* magazine, accessed May 30, 2021, https://shiftermagazine.com /about.

microaggressions, individual content creators can still play a role in ameliorating toxic online climates.

Building from Pierce's work on *Sesame Street*, Rini (2021) provides another potential solution to the problem of environmental microaggressions: content curation. Rini shares Pierce's concern that media can perpetuate bias in its audience, but whereas Pierce focused on the duties of content *creators*, Rini suggests that individual *consumers* of media could also play a role in mitigating the effects of environmental microaggressions. She argues that individuals can exert "remote control" over the biases they develop by intentionally limiting their consumption of stereotypical media and seeking out counterstereotypical portrayals and shows that feature diverse casts and writers.[27] Curating the content we consume allows us to shift our habits of stereotypical thought and, ultimately, reduce our propensity to commit microaggressions.

Turning back to the online context, individual search engine users could adopt a similar strategy of content curation by changing what kind of queries they search for. Instead of seeking out stereotypical content, search engine users could look for more accurate portrayals of marginalized groups. If enough people typed "Why are Black women not being paid equally" or "Why are Black men the most likely to die at the hands of the police," then Google wouldn't need to carefully monitor (or censure) its autocomplete suggestions for those groups. Good searches would create good autocomplete predictions, which would in turn inspire further good searches—a virtuous cycle.

Both content creation and content curation could be rendered more effective with institutional support. Like Pierce's *Sesame Street*, which was displayed on the Public Broadcasting Service, counterstereotypical online content could achieve greater popularity if it had government funding and advertising. Governments could also invest in educational initiatives that would provide online literacy and guidance about how to seek out more accurate information about marginalized groups. If more of the primary and secondary school curriculum was devoted to countering stereotypes and teaching the truth about historical and current systems of oppression, search engine users would have a better starting point for personal research and content curation.

All the approaches we've surveyed thus far could reduce the prevalence of stereotyping autocomplete microaggressions without any changes to Google's algorithm. This is the kind of solution Google (and other tech corporations) has advocated for all along: since algorithms merely mirror the bias of society's searches, if we reduce bias in society, *mutatis mutandis* we'll reduce algorithmic

[27] See also Dean, Victor, and Guidry-Grimes (2016) for a similar approach to disrupting microaggressive bias at the individual level.

microaggressions as well. As we showed in section 2, however, when it comes to algorithms, claims of neutrality are misleading. Google's autocomplete algorithm doesn't just mirror society's biases—it perpetuates bias and perpetrates microaggressions. Therefore, in addition to efforts by individuals and institutions, Google has a duty to fix the problems it helped create. We'll conclude by outlining the unique role Google can play in reducing algorithmic microaggressions, which we hope will be instructive for other corporations facing similar issues. Instead of the reactive strategies that have been tried thus far (enumerated in section 3), our recommendations will require Google to be proactive: retraining its algorithm and, when necessary, restraining it. We'll explain each of these approaches, below.

First, retraining. Google's autocomplete is receiving constant inputs that over time can lead it to slightly update its suggestions. However, its first incarnation was created in 2004, and like other Google algorithms, it was probably trained on a dataset from an even earlier era (Garber 2013). Levendowski (2018) has critiqued the practice of training algorithms on biased data—often data old enough to be in the public domain or, worse, made public during a criminal investigation. (As we've mentioned, many NLP algorithms were trained on a dataset that included the Enron emails, which are predictably filled with racism and sexism, in addition to other wrongdoing.) We would thus join Levendowski in calling for Google and other tech companies to commit to retraining their algorithms on better, less biased datasets.

Moreover, we'd further encourage tech companies to take a page from Pierce's book. Pierce knew the value of education, and we can apply his ideas about the education of children to the education of algorithms. Rather than being satisfied with preexisting datasets, Google could invest in creating their own, more inclusive datasets that would train their algorithms with a "hidden curriculum" of social equality. To give just one example, Google (and other tech giants) could invest in digitizing archives of current and historical activist movements. Activist archives are often precariously funded, if their digitization is funded at all, so if Google offered to host these archives for free, it could benefit all parties. Imagine if NLP algorithms were trained on archives of Black Lives Matter emails or LGBTQ2S oral histories, instead of Enron![28] This method would work best for NLP trained on small datasets. Its effectiveness would diminish as the size of the training dataset increases—for instance, the impact on a dataset the size of GPT-3, in which all of Wikipedia is only 3 percent of the total dataset, would be very small—but progress is being made on machine learning using smaller datasets. Chahal and Toner (2021) have recently argued for further attention to small data techniques like transfer learning, where an

[28] Obviously, there would need to be protections put in place, including but by no means limited to informed consent specifying that the data will be used solely for training purposes (and not, for example, shared with law enforcement agencies).

algorithm previously primed on a large dataset could be effectively trained on a much smaller novel dataset. We would encourage Google, and other companies, to invest in these retraining technologies that could employ pared down datasets—in this case, datasets with a greater proportion of content created by members of marginalized groups (and other content that avoids reinforcing stereotypes). The current datasets teach algorithms to commit microaggressions, whereas more representative datasets could teach them to reflect a better world of social equality.

Retraining algorithms and investing in tech solutions could help, but innovation will take time, as more representative content is sought out, created, and digitized. Moreover, no dataset can perfectly train an algorithm to avoid microaggressions, as new microaggressions are constantly being identified. So what should Google do when algorithmic microaggressions are brought to their attention? Here is where our second recommendation becomes relevant. Until more comprehensive retraining is possible, the autocomplete algorithm should be restrained. The algorithm should be turned off for searches about marginalized groups, but instead of showing no results, the autocomplete should be populated with a list of suggestions curated by humans—namely, people with lived and research expertise in recognizing microaggressions and combating oppression.

This intervention would not be as much of a departure as it might seem. From its inception, in addition to being a time-saving tool, autocomplete was also envisioned as a tool that would allow users to "learn about things you haven't dreamt of."[29] We suggest embracing this liberatory potential inherent in exposing users to content they might never otherwise have dreamt of. What if Google autocomplete were not merely a mirror of current biases but instead a nudge towards justice, an educational tool that suggests the questions you should be asking? As we've shown, there is no neutrality—in choosing not to intervene, Google is already adopting a morally valenced stance and perpetuating microaggressions—so it should choose to embrace the power of suggestion. In addition to teaching us facts like the boiling point of water, Google autocomplete could teach us how to best complete the sentence, "Why are Black women . . ."

- ". . . dying in childbirth?"
- ". . . known for their contributions to social justice?"
- ". . . underrepresented in business?"
- ". . . at the forefront of vaccine research?"
- ". . . the victims of misogynoir?"
- ". . . only famous for achievements in entertainment and sport?"
- ". . . stereotyped in film?"

[29] The creator, Kevin Gibbs, quoted in Gerber (2013).

- ". . . the founders of Black Lives Matter?"
- ". . . gaining more recognition?"
- ". . . researching intersectionality?"

Of course, this list of counterstereotypical exemplars and accurate portrayals of oppression is just a starting point. Instead of taking our word for it, Google should gather—and pay—a team of experts (including those with expertise gained from lived experience) to develop the best possible list and to keep that list continually updated to reflect the changing social world.

How might such a change in institutional policy actually occur? We hope that a change like this could be motivated internally by Google itself or its employees; after all, the employees have successfully pressured Google to change its policies in several high-profile instances. Perhaps more likely, though, is that it could be driven by external pressure, either legislative or social. We might look to Facebook's decision to establish an oversight board to review its decisions over what content to allow and what to block on its site.[30] In the face of public pressure over its decisions, Facebook preferred to create an independent body that would have final say. We can imagine similar social pressure on Google that would make them prefer to defer some decisions from algorithms to arm's-length independent entities; this could help address algorithmic microaggressions while still allowing Google to remain insulated from any (unpopular) decisions made by these entities. Alternatively, new legislation could force Google to adopt more socially responsible policies with respect to its algorithms. This could be accomplished directly through new regulation, or indirectly by removing existing exemptions from civil, collective action lawsuits (like Section 230) and forcing Google to pay for the harm it has caused to members of marginalized groups.[31]

5. Conclusion

Google and other tech companies have assumed that they should remain neutral on issues of social justice, but we've shown that algorithms are not morally inert. Doing nothing is not an option, yet Google's current policy of removing unpopular autocompletes is both ad hoc and damaging to the very groups it's ostensibly trying to protect. Removing stereotyping microaggressions creates an erasure microaggression that impedes marginalized users from accessing knowledge and combating internalized stereotypes. We've argued that to truly serve the needs of its marginalized users, Google should embrace the liberatory power of suggestion.

[30] See Klonick (2020) and Douek (2019) for more on the Facebook oversight board and its implications.

[31] Though see Morrison (2020) for a discussion of the pros and cons of Section 230.

Rather than mirroring society's bias, Google autocomplete could be a force for good and a guiding light for other tech companies who want to avoid algorithmic microaggressions.

References

Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes." *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November): 1–30. https://doi.org/10.1145/3359301.

Allen, Antoine. 2016. "The 'Three Black Teenagers' Search Shows It Is Society, Not Google, That Is Racist." *Guardian*, June 10, 2016. https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Basu, Rima. 2019. "Radical Moral Encroachment: The Moral Stakes of Racist Beliefs." *Philosophical Issues* 29, no. 1 (October): 9–23. https://doi.org/10.1111/phis.12137.

Benjamin, Ruha. 2019. "Coded Exposure: Is Visibility a Trap?" In *Race After Technology: Abolitionist Tools for the New Jim Code*, ch. 3. Cambridge: Polity Press.

Brennan, Samantha. 2013. "Rethinking the Moral Significance of Micro-Inequities: The Case of Women in Philosophy." In *Women in Philosophy: What Needs to Change?*, edited by Katrina Hutchinson and Fiona Jenkins, 180–96. Oxford: Oxford University Press.

———. 2016. "The Moral Status of Micro-Inequities: In Favor of Institutional Solutions." In *Implicit Bias and Philosophy, Vol. 2: Moral Responsibility, Structural Injustice, and Ethics*, edited by Michael Brownstein and Jennifer Saul, 235–53. Oxford: Oxford University Press.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In "Conference on Fairness, Accountability and Transparency," Proceedings of Machine Learning Research 81:77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Cadwalladr, Carole. 2016. "Google, Democracy and the Truth about Internet Search." *Guardian*, December 4, 2016. https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science* 356, no. 6334 (April 14): 183–86. https://doi.org/10.1126/science.aal4230.

Chahal, Husanjot, and Helen Toner. 2021. "'Small Data' Are Also Crucial for Machine Learning." *Scientific American*, October 19, 2021. https://www.scientificamerican.com/article/small-data-are-also-crucial-for-machine-learning/.

Chowdhury, Rumman. 2021. "Sharing Learnings about Our Image Cropping Algorithm." Twitter "Insights" blog, May 19, 2021. https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

Corbett, Elisha. 2019. "When Disinformation Becomes Deadly: The Case of Missing and Murdered Indigenous Women and Girls in Canadian Media." *Disinformation and Digital Democracies in the 21st Century*, edited by Joseph McQuade, Tiffany Kwok, and James Cho, 19–23. Toronto: NATO Association of Canada.

Danaher, John. 2016. "Robots, Law and the Retribution Gap." *Ethics and Information Technology* 18, no. 4 (December): 299–309. https://doi.org/10.1007/s10676-016-9403-3.

Dean, Megan A., Elizabeth Victor, and Laura Guidry-Grimes. 2016. "Inhospitable Healthcare Spaces: Why Diversity Training on LGBTQIA Issues Is Not Enough." *Bioethical Inquiry* 13, no. 4 (December): 557–70. https://doi.org/10.1007/s11673-016-9738-9

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism.* Cambridge, MA: MIT Press.

DiversityInc. 2016. "9 Things NOT to Say to Women Coworkers." *DiversityInc*, February 19, 2016. https://www.diversityinc.com/things-not-to-say-to-women-coworkers/.

Dotson, Kristie. 2011. "Tracking Epistemic Violence, Tracking Practices of Silencing." *Hypatia* 26, no. 2 (Spring): 236–57. https://doi.org/10.1111/j.1527-2001.2011.01177.x.

Douek, Evelyn. 2019. "Facebook's Oversight Board: Move Fast with Stable Infrastructure and Humility." *North Carolina Journal of Law & Technology* 21 (1). https://ncjolt.org/articles/6436-2/.

Fatima, Saba. 2017. "On the Edge of Knowing: Microaggression and Epistemic Uncertainty as a Woman of Color." In *Surviving Sexism in Academia: Feminist Strategies for Leadership*, edited by Kirsti Cole and Holly Hassel, 147–54. New York: Routledge.

Freeman, Lauren, and Heather Stewart. 2018. "Microaggressions in Clinical Medicine." *Kennedy Institute of Ethics Journal* 28, no. 4 (December): 411–49. https://doi.org/10.1353/ken.2018.0024.

Fricker, Miranda. 2007. “Hermeneutical Injustice.” In *Epistemic Injustice: Power and the Ethics of Knowing*, ch. 7. Oxford: Oxford University Press.

Friedlaender, Christina. 2018. “On Microaggressions: Cumulative Harm and Individual Responsibility.” *Hypatia* 33, no. 1 (Winter): 5–21. https://doi.org/10.1111/hypa.12390.

———. 2021. “Putting Non-Binary People into Binary-Driven Design: Environmental Microaggressions and Existential Erasure in Digital Life.” Paper presented at APA Committee on LGBTQ Issues in the Profession American Philosophical Association, Central Division. February 2021.

Friedlaender, Christina, and Veronica Ivy. 2020. “A Defense of Intentional Microaggressions and Microaggressive Harassment: The Fundamental Attribution Error, Harassment, and Gaslighting of Transgender Athletes.” In *Microaggressions and Philosophy*, edited by Lauren Freeman and Jeanine Weekes Schroer, 184–204. New York: Routledge.

Frye, Marilyn. 1983. “Oppression.” In *Politics of Reality*, 1–16. Berkely, CA: Crossing Press.

Garber, Megan. 2013. “How Google’s Autocomplete Was . . . Created/Invented/Born.” *Atlantic*, August 23, 2013. https://www.theatlantic.com/technology/archive/2013/08/how-googles-autocomplete-was-created-invented-born/278991/.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Nets.” *Advances in Neural Information Processing Systems* 27, edited by: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 2672–80.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. “Explaining and Harnessing Adversarial Examples.” arXiv preprint arXiv:1412.6572 [stat.ML]. https://doi.org/10.48550/arXiv.1412.6572.

Google. 2021. “Autocomplete Policies.” https://support.google.com/websearch/answer/7368877?hl=en.

Greene, Bryan. 2019. “The Unmistakable Black Roots of ‘Sesame Street’.” *Smithsonian Magazine*, November 7, 2019. https://www.smithsonianmag.com/history/unmistakable-black-roots-sesame-street-180973490/.

Harrington, Anne. 2019. “Psychiatry, Racism, and the Birth of ‘Sesame Street.’” *Undark*, May 17, 2019. https://undark.org/2019/05/17/psychiatry-racism-sesame-street/.

Henning, Tempest M. 2020. “Racial Methodological Microaggressions: When Good Intersectionality Goes Bad.” In *Microaggressions and Philosophy*, edited by Lauren Freeman and Jeanine Weekes Schroer, 251–72. New York: Routledge.

Hern, Alex. 2020. “Twitter Apologises for ‘Racist’ Image-Cropping Algorithm.” *Guardian*, September 21, 202. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm.

Himmelreich, Johannes. 2018. “Agency and Embodiment: Groups, Human–Machine Interactions, and Virtual Realities.” *Ratio* 31, no. 2 (June): 197–213. https://doi.org/10.1111/rati.12158.

Jain, Niharika, Lydia Manikonda, Alberto Olmo Hernandez, Sailik Sengupta, and Subbarao Kambhampati. 2018. “Imagining an Engineer: On GAN-Based Data Augmentation Perpetuating Biases.” arXiv preprint arXiv:1811.03751 [cs.LG]. https://doi.org/10.48550/arXiv.1811.03751.

Johnson, Gabbrielle. Forthcoming. “Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning.” *Journal of Moral Philosophy*.

Klonick, Kate. 2020. “The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression.” *Yale Law Journal* 129, no. 8 (June): 2418–99. https://www.yalelawjournal.org/feature/the-facebook-oversight-board.

Köhler, Sebastian. 2020. “Instrumental Robots.” *Science and Engineering Ethics* 26, no. 6 (December): 3121–41. https://doi.org/10.1007/s11948-020-00259-5.

Lapowsky, Issie. 2018. “Google Autocomplete Still Makes Vile Suggestions.” *Wired*, February 12, 2018. https://www.wired.com/story/google-autocomplete-vile-suggestions/.

Levendowski, Amanda. 2018. “How Copyright Law Can Fix Artificial Intelligence’s Implicit Bias Problem.” *Washington Law Review* 93, no. 2 (June): 579–630. https://digitalcommons.law.uw.edu/wlr/vol93/iss2/2/.

Li, Chuan. 2020. “OpenAI’s GPT-3 Language Model: A Technical Overview.” *Lambda Deep Learning Blog*, June 3, 2020. https://lambdalabs.com/blog/demystifying-gpt-3/.

Longino, Helen E. 1995. “Gender, Politics, and the Theoretical Virtues.” *Synthese* 104, no. 3 (September): 383–97. https://doi.org/10.1007/BF01064506.

Mahesh, Batta. 2020. “Machine Learning Algorithms—A Review.” *International Journal of Science and Research (IJSR)* 9, no. 1 (January): 381–86.

Matthias, Andreas. 2004. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6, no. 3 (September): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McClure, Emma. 2020. “Escalating Linguistic Violence: From Microaggressions to Hate Speech.” In *Microaggressions and Philosophy*, edited by Lauren Freeman and Jeanine Weekes Schroer, 121–45. New York: Routledge.

McClure, Emma, and Regina Rini. 2020. “Microaggression: Conceptual and Scientific Issues.” *Philosophy Compass* 15, no. 4 (April): e12659. https://doi.org/10.1111/phc3.12659.

McTernan, Emily. 2018. "Microaggressions, Equality, and Social Practices." *Journal of Political Philosophy* 26, no. 3 (September): 261–81. https://doi.org/10.1111/jopp.12150.

Metz, Cade. 2016. "How Google's AI Viewed the Move No Human Could Understand." *Wired,* March 14, 2016. https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/.

Metz, Cade and Wakabayashi, Daisuke. 2020. "Google Researcher Says She Was Fired over Paper Highlighting Bias in A.I." *New York Times*, December 3, 2020. https://www.nytimes.com/2020/12/03/technology/google-researcher-tim nit-gebru.html.

Millar, Jason. 2015. "Technology as Moral Proxy: Autonomy and Paternalism by Design." *IEEE technology and Society Magazine* 34, no. 2 (June): 47–55. https://doi.org/10.1109/MTS.2015.2425612.

Mills, Charles. 2007. "White Ignorance." In *Race and Epistemologies of Ignorance*, edited by Shannon Sullivan and Nancy Tuana, 13–38. Albany, New York: SUNY Press.

Morrison, Sara. 2020 "Section 230, the Internet Free Speech Law Trump Wants to Repeal, Explained." *Vox*, October 6, 2020. https://www.vox.com/recode/2020/5/28/21273241/section-230-explained-trump-social-media-twitter-facebook.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press

Nyholm, Sven. 2018. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24, no. 4 (August): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.

Osareh, Alireza, and Bita Shadgar. 2010. "Machine Learning Techniques to Diagnose Breast Cancer." *2010 5th International Symposium on Health Informatics and Bioinformatics*, edited by Nazife Baykal, Osman Saka, Mesude İşcan, and Tolga Can 114–20. Ankara: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HIBIT.2010.5478895.

Patil, Swati P., B. V. Pawar, and Ajay S. Patil. 2013. "Search Engine Optimization: A Study." *Research Journal of Computer and Information Technology Sciences* 1, no. 1 (February): 10–13.

Pérez Huber, Lindsay, and Daniel G. Solórzano. 2015. "Racial Microaggressions as a Tool for Critical Race Research." *Race Ethnicity and Education* 18 (3): 297–320. https://doi.org/10.1080/13613324.2014.994173.

Pierce, Chester. 1970. "Offensive Mechanisms." In *The Black Seventies*, edited by Floyd B. Barbour, 265–82. Boston: Porter Sargent.

Pierce, Chester M. (1974) 2015 "Psychiatric Problems of the Black Minority." In *American Handbook of Psychiatry*, edited by Silvano Arieti, Vol. 2, *Child and*

Adolescent Psychology, Sociocultural and Community Psychiatry, edited by Gerald Caplan, ch. 33. New York: Basic Books.

Pohlhaus, Gaile, Jr. 2012. "Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance." *Hypatia* 27, no. 4 (November): 715–35. https://doi.org/10.1111/j.1527-2001.2011.01222.x.

Reiheld, Allison. 2020. "Microaggressions as a Disciplinary Technique for Fat and Potentially Fat Bodies." In *Microaggressions and Philosophy*, edited by Lauren Freeman and Jeanine Weekes Schroer, 205–25. New York: Routledge.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier." In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. New York: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rini, Regina. 2021. *The Ethics of Microaggressions.* New York: Routledge.

Roy, Senjooti, and Liat Ayalon. 2020. "Age and Gender Stereotypes Reflected in Google's 'Autocomplete' Function: The Portrayal and Possible Spread of Societal Stereotypes." *The Gerontologist* 60, no. 6 (September): 1020–28. https://doi.org/10.1093/geront/gnz172.

Smith, Aaron. 2018. *Public Attitudes toward Computer Algorithms*. Pew Research Center, November 16, 2018. https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (February): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Steinfeldt, Jesse A., Jacqueline Hyman, and M. Clint Steinfeldt. 2019. "Environmental Microaggressions: Context, Symbols, and Mascots." In *Microaggression Theory: Influence and Implications*, edited by Gina C. Torino, David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue, 213–25. New York: Wiley. https://doi.org/10.1002/9781119466642.ch13.

Sue, Derald Wing. 2010. *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation.* New York: Wiley.

Thoma, Johanna. 2022. "Risk Imposition by Artificial Agents: The Moral Proxy Problem." In *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, edited by Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard, 50–66. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009207898.006.

Torino, Gina C., David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue. 2019. "Microaggression Theory: What the Future Holds." In *Microaggression Theory: Influence and Implications,* edited by Gina C. Torino, David P. Rivera, Christina M. Capodilupo, Kevin L. Nadal, and Derald Wing Sue, 309–28. New York: Wiley. https://doi.org/10.1002/9781119466642.ch19.

Tschaepe, Mark. 2021. "Pragmatic Ethics for Generative Adversarial Networks: Coupling, Cyborgs, and Machine Learning." *Contemporary Pragmatism* 18, no. 1 (May): 95–111. https://doi.org/10.1163/18758185-bja10005.

Yee, Kyra, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. "Image Cropping on Twitter: Fairness Metrics, Their Limitations, and the Importance of Representation, Design, and Agency." arXiv preprint arXiv:2105.08667 [cs.CY]. https://doi.org/10.48550/arXiv.2105.08667.

EMMA McCLURE (she/they) is an assistant professor of philosophy at Saint Mary's University (Halifax) working at the intersection of ethics, feminism, philosophy of law, and critical race theory. Their research focuses on various topics within the ethics of conversation—ranging from moral responsibility for microaggressions to supporting trauma recovery and self-reintegration.

BENJAMIN WALD (he/him) is a research analyst at the Chiefs of Ontario. He wrote the bulk of this paper while a postdoctoral fellow at the Schwartz Reisman Institute for Technology and Society. His research interests include AI ethics, indigenous data sovereignty, and the relationship between ethics and agency.