Reflecting on Bias

In my classes, students are required to complete a series of reflection exercises. In each exercise, they read one or more articles and write a reflection that addresses a series of prompts. This is usually followed by an in-class discussion. The class deals with a wide variety of ethical and social issues surrounding AI and technology; this submission focuses specifically on bias.

This specific set of assignments provides students with the opportunity to read about and reflect on the many different ways in which bias can occur in AI and ML problems. Students will learn about the technical definition of bias, how this relates to the more colloquial version, why it is so difficult to define what it means to be fair, and the importance of understanding what biases exist in real-world data sets. Also included are strategies and resources for class discussion and a discussion of specifications grading as an approach for assigning written work. These assignments have been used in a variety of classes, including both first-year programming and Master's level AI courses, to help students think about the impact of what they build and the ways in which technology can either disrupt or perpetuate biases.

This module consists of four reflections on different aspects of bias. These exercises have several goals:

1. To help students understand the real-world impact of the software they design, and to teach them to think critically about these impacts.
2. To connect the technical knowledge they are learning to real-world problems.
3. To deepen their understanding of the different forms of bias that can occur.

Specifications Grading

A common concern with written assignments in computer science courses is grading, and everything associated with it. I'm not a rhetoric instructor, and my students, especially those for whom English is not their first language, are often concerned about their writing and grammar skills. These issues can lead to a lot of student anxiety.

The solution I've adopted for my courses is to use specifications grading.
[Nilson:15]. Specifications grading encourages the instructor to provide a description of what is (and is not) expected for the assignment; those assignments that meet the specifications receive full credit. In the case of the reflections, my instructions to the students are:

1. You need to address all of the question prompts.
2. I need to be able to understand what you are trying to say.
3. You need to back up claims with evidence, and cite appropriately.

Students who do this receive full credit. I give half credit for incomplete submissions, and no credit for missing work.

Please note that I do not expect perfect grammar, and I don't provide feedback on writing style or ESL issues. I also don't expect all students to agree with me, or the articles, but I do expect them to think things through.

Follow-up

Once the students have submitted their reflections, we typically discuss the issue in class. We will often break into pairs or small groups for initial discussion, in order to create opportunities for quieter students to participate and share ideas. I then ask the groups to summarize or report out the discussion for the class. Some topics that I try to look for and emphasize are:

- Different measures of fairness produce different outcomes. Which of those outcomes is most preferable is not obvious; students see in the diversity exercise that different ranking metrics can produce qualitatively different results, and in the COMPAS example they see that it's not possible to have a uniform risk score that treats defendants of different races equally.
- A learning algorithm will reflect the biases in its data. We do a lot of technical discussion of bias and sampling in class; the COMPAS and Quartz pieces provide clear, real-world illustrations of the problems that an ML approach can run into when using real-world data sets that contain bias.
- For some students, this has provided a clearer understanding of the difference between being "not racist" and being "antiracist." As we start to look more carefully at the COMPAS problem in particular, the question "why are Black parolees re-arrested at higher rates?" comes to the fore, as this is at the core of COMPAS' errors. When students are asked "how would we as computer scientists address this problem," they begin to see that it's not just a matter of not using race as a criterion, but rather of actively taking steps to undo a systemic bias.
- System designers can inadvertently introduce bias; we need to be aware of this as we design applications. The Bias on the Web article illustrates this with several scenarios. I sometimes follow this with a discussion of user-centered design, and the need to engage end users in software development early on in order to better understand systemic effects of design decisions.

While I have not gathered assessment data to evaluate the effectiveness of these exercises, anecdotally I have been quite pleased with both the feedback I've received and the quality of the student work. Many students have told me that they appreciated the ability to more deeply technical knowledge to broader impact, and to better understand the relationship between the AI and ML concepts we learn about in class and the issues they hear about in the news. As an educator, I feel that it's my responsibility to help the students who will be designing tomorrow's ML systems be aware of, and hopefully avoid, some of the problems experienced by today's systems.

References:

Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time. Nilson, Linda B. United States: Stylus Publishing.