Equity Beyond Bias in Language Technologies for Education

Elijah Mayfield[1]*, **Michael Madaio**[2], **Shrimai Prabhumoye**[1], **David Gerritsen**[3], **Brittany McLaughlin**[4], **Ezekiel Dixon-Román**[5], and **Alan W Black**[1]
[1]Language Technologies Institute, [2]Human-Computer Interaction Institute, and
[3]Eberly Center for Teaching Excellence, Carnegie Mellon University;
[4]ScribbleUp Learning; [5]School of Social Policy and Practice, University of Pennsylvania
* Corresponding author: elijah@cmu.edu

Abstract

There is a long record of research on equity in schools. As machine learning researchers begin to study fairness and bias in earnest, language technologies in education have an unusually strong theoretical and applied foundation to build on. Here, we introduce concepts from culturally relevant pedagogy and other frameworks for teaching and learning, and identify future work on equity in NLP. We present case studies in a range of topics like intelligent tutoring systems, computer-assisted language learning, automated essay scoring, and sentiment analysis in classrooms, and provide an actionable agenda for research.

1 Introduction

Researchers across machine learning applications are finding unintended outcomes from their systems, with inequitable or even unethical impacts (Barocas and Selbst, 2016). We are at an inflection point in the study of fair machine learning; popular science publications are shedding light on the widespread impacts of algorithmic bias (Noble, 2018; Eubanks, 2018; Angwin et al., 2016) and specialized technical conferences like ACM FAT*[1] and FATML[2] now provide methods and examples of research addressing ethics in model bias, the design of datasets, and user interfaces for algorithmic interventions. "Impact" investing in educational technology[3] has grown (Gates Foundation and Chan Zuckerberg Initiative, 2019) and these machine learning tools are now pervasive in educational decision-making (Wan, 2019). Yet in recent literature reviews of NLP in edtech, the focus has been on narrowly scoped technical topics, like speech (Eskenazi, 2009) or text and chat data (Litman, 2016), but crucially, do not address equity issues more broadly. NLP applications are mainstays in schools and have great reach, a trend poised to accelerate with the adoption of interactive, language-enabled devices like Alexa, both at home and in the classroom (Ziegeldorf et al., 2014; Horn, 2018; Boccella, 2019). As a field, we risk unwittingly contributing to harm for learners if we don't understand the ethical consequences of our research – but we don't have to start from scratch.

Education philosophers have long advocated for equity in schooling for all learners (Dewey, 1923; Freire, 1970), and over decades, have built rich pedagogies to accomplish goals of social justice for students (Ladson-Billings, 1995); this work has flourished in progressive schools (Morrell, 2015; Paris and Alim, 2017). Developers of edtech have already moved from technological innovation for its own sake, to a focus on efficacy and learning analytics, tying educational data mining to specific student outcomes (Baker and Inventado, 2014). This paper presents a roadmap for now incorporating equity into the design, evaluation, and implementation of those systems.

In sections 2 and 3 we give overviews of existing research, first on fair machine learning, then on social justice pedagogies in education. The bulk of our new contributions are in section 4-7, where we describe key problem areas for NLP researchers in education. We conclude with practical recommendations in section 8.

2 Primer on Fair Machine Learning

The topic of ethics in technology dates back to decades ago (Winner, 1989); but uptake of conversations about building equitable algorithmic systems is fairly recent. The existing literature prioritizes topics of bias and fairness, mostly based on what some have called *"allocational harm"* (Crawford, 2017). Researchers measure the distribution of outcomes produced by automated decision-making, and evaluate whether subgroups received proportional shares of a resource being

[1]https://fatconference.org/2019/
[2]http://www.fatml.org/
[3]From this point forward, abbreviated as "edtech."

distributed - bail release recommendations, approval for a mortgage, high test scores, and so on. Over and over, differential outcomes have been tied to biased modeling along demographic lines like gender, race, and age (Friedler et al., 2019).

Some have questioned the value of fairness research, arguing that machine learning may simply reproduce existing distributions, rather than cause harm in itself (see Mittelstadt et al. (2016) for an overview of this debate). But high-profile research has repeatedly shown an *amplifying* effect of machine learning on concrete real-world outcomes, like racial bias in recidivism prediction in judicial hearings (Corbett-Davies et al., 2017), or disproportionate error from facial recognition for dark skin tones, particularly among individuals identifying as female (Buolamwini and Gebru, 2018).

Fairness work in NLP has focused particularly in dense semantic representations at the lexical or sentence level. In learned embeddings of meaning, bias exists along race and gender lines (Caliskan et al., 2017; Garg et al., 2018) and is passed downstream, producing biased outcomes for tasks like coreference resolution (Zhao et al., 2018a), sentiment analysis (Kiritchenko and Mohammad, 2018), search (Romanov et al., 2019), and dialogue systems (Voigt et al., 2018; Henderson et al., 2018). Research beyond metrics, analyzing the broader social impact of biased NLP, has also begun (Hovy and Spruit, 2016).

Many of these problems stem from training data selection; models trained on standard written professional English, like the Penn Treebank (Marcus et al., 1993), fail to transfer to other writing styles, especially online where research suggests that NLP performance is degraded for underrepresented language groups, like African American English (Petrov and McDonald, 2012; Blodgett et al., 2017). Early work on "de-biasing" NLP has begun, seeking to reduce the amplification of bias in dense word embeddings (Bolukbasi et al., 2016; Zhao et al., 2017, 2018b); but early results still leave room for improvement (Gonen and Goldberg, 2019). Accounting for dialects and other language variation has been moderately more successful, with examples in speech recognition (Kraljic et al., 2008), parsing (Gimpel et al., 2011), and classification (Jurgens et al., 2017).

There are many open questions. Chouldechova (2017) and Corbett-Davies and Goel (2018) work to even *define* fairness, giving several proposals; but related research has shown these definitions are brittle. Classifiers may trivially fail to maintain fairness properties when the output from one classifier is used as input for another, for instance (Dwork and Ilvento, 2018), or even worsen disparate outcomes after iterating on algorithmic predictions over time (Liu et al., 2018). Research in computational ethics (Hooker and Kim, 2018) may give some guidelines for the NLP community broadly, and work on richer formal systems of guarantees on fairness is underway (Kearns et al., 2019); but while this research is ongoing, developers continue to build systems. For NLP researchers working in education, specifically, a key resource will be the long tradition of educational equity research and praxis that exists today, and is being practiced in schools already.

3 Equity in Education Research

Machine learning research in general tends to focus on recent publication; to counteract this and set a longer-term context, in the following section we explain the historical background on learning science research that considers socio-cultural dimensions of learning and their implications for equity, work that motivates our recommendations for technologists building educational interventions.

3.1 Sociocultural and Critical Perspectives

While much of the earliest work on learning science was purely behaviorist, the field's expansion into sociocultural factors that affect learning is old, beginning nearly a century ago. Driven by Marxist philosopher and psychologist Lev Vygotsky, the gaze of research shifted from inner processes of the mind to interactions between students and their cultural context and practices (Moll, 1992). This tradition drove research into individual development via socially-mediated processes of learning (Chaiklin, 2003). The mediated learning experience is done via a process of scaffolding what the learner knows and what they need help on, in their "zone of proximal development" (Hammond and Gibbons, 2005). This work also acknowledged the connection between formal school education and informal education in the world (Scribner and Cole, 1973), and introduced the idea of learning as a social process in which students build identity (Wenger, 2010). This conceptual framework now dominate the scientific discourse on sociocultural research in edtech systems (Aleven et al., 2016).

The sociocultural paradigm from Vygotsky has humanized education compared to purely behaviorist approaches; meanwhile, parallel work in the emerging field of *critical pedagogy* was taking more aggressive steps. Led by Brazilian educational and social philosopher Freire (1970), this work argued that formal schooling was an ideological system for preserving existing power structures, that treats students as receptacles to be filled with culturally dominant views (*i.e.* a "banking" model), rather than giving students the opportunity to learn topics of intrinsic meaning to them. This alternate approach led to unprecedented gains in adult literacy during the twentieth century, particularly in Brazil (Kirkendall, 2010) and in Cuba (Samuel and Williams, 2016), demonstrating what pedagogical theorists described as liberatory education and critical consciousness (Freire, 1985). This, and later work by critical theorists like hooks (2003), critiqued the banking model where learning is viewed as providing neutral information to students. Critical pedagogy instead views teaching as a fundamentally political process, where students may engage with topics from their life, ask questions about their contexts, and identify systemic power relations and institutions. When applied in school contexts, this approach successfully reaches students typically left behind in more mainstream pedagogies (Morrell, 2015).

Multiculturalist approaches to education build on this, drawing from cultural, ethnic, and womens studies to teach by drawing on students' own cultural history and practices. The goal is to promote equity through learning within a student's community and culture, producing a *culturally sustaining pedagogy* (Ladson-Billings, 1995). This approach necessitates educators who come from, or are deeply competent in, the cultural norms and expressions of their students, creating content and opportunities that allow students to connect with learning in an affirming way. By giving students tools to engage with and critique society, the most recent approaches continue to enable student growth (Paris and Alim, 2017).

3.2 Application to Algorithms in Edtech

These perspectives can be hard to align with technological interventions. As direct critiques of dominant ideologies and institutions that legitimate and maintain inequality for students, their language is more forceful than most machine learning research. Unlike fairness literature in computer science venues, these works explicitly describe existing practices as based in white supremacist patriarchy, heteronormativity, and colonialism. This makes these pedagogies more expressive, capable of defining a path forward for equitable technologies; but it also makes them more suspicious of interventions that scale without local context and cultural knowledge.

However, educators have successfully applied these principles in technology-oriented work. Mislevy et al. (2009) shows how critical analysis can support and define assessment; Morris and Stommel (2018) uses them to develop a digital pedagogy. The Gordon Commission shows how critical work can be a basis for development of adaptive learning systems (Armour-Thomas and Gordon, 2013). Across these and other applications, some principles are immediately clear:

- A shift in the goal of assessment, from measuring *static knowledge* to assessing *formative process*, acknowledging student growth at least as much as facts they have "banked."
- A vocabulary and willingness to describe existing systems as oppressive for students, on lines of race, economic class, gender, physical abilities, and other aspects of identity.
- A demand for cultural competence from the teachers and designers of learning systems, aligning the creators of educational environments with the students they teach.

The remainder of our paper summarizes key recommendations that lead from these principles. We reference them in the hope that researchers will move their conversations about equity in machine learning beyond model bias and allocational harm for subgroups. Such work is vital and the task of bias measurement is not solved yet, but researchers are already racing to build tools for these problems. Madnani et al. (2017), for instance, presents a capable tool for evaluating fair outcomes in automated essay scoring. It would be a mistake to focus on bias alone Given existing pedagogical work on equity and its focus on learning through dialogue, critical discourse, and action, we can propose broader mindset shifts for researchers. Our goal is to avoid harm to students and prevent expenditure of resources on research that maintains inequity rather than closing gaps in achievement across student populations.

4 Avoiding Representational Harms

First, beyond allocational harm, there are *"representational harms"* in machine learning (Crawford, 2017). This class of issues includes the ways in which technologies represent groups of people or cultures. This may take the form of search results returning stereotypical images of minorities (Noble, 2018) or other algorithmic stereotyping (Abbasi et al., 2019); much of the work in word embeddings falls into this category (Caliskan et al., 2017), though research on downstream tasks and outcomes often have more allocational focus. Machine learning may also marginalize groups by simply *not* representing their culture, resulting in educational systems where learners do not see themselves in the texts selected by instructors.

These harms can exist even when no disparate outcomes are observed, and even if there is no measured gap in predictive accuracy of models (Binns, 2018). Students whose cultural background is in the minority in a classroom are less prone to participate in teacher-student interactions (Tatum et al., 2013) and in student group discussion (White, 2011); these variations are predictable by gender, race, and nationality (Eddy et al., 2015). We also know that instructor credibility is tied to demographics (Bavishi et al., 2010), as are student evaluations of a teacher's trustworthiness and caring (Finn et al., 2009).

4.1 Case Study: Agent-based Intelligent Tutoring Systems

In intelligent tutoring systems (ITS), a human-like agent or visual avatar engages with students through text or speech. These systems now pair natural language instruction with parasocial features (Lubold et al., 2018) and mimicking nuanced human behaviors like finding "teachable moments" (Nye et al., 2014). They are used individually or with groups of students (Kumar et al., 2007) and to provide narrowly targeted support for Autistic students (Nojavanasghari et al., 2017) and deaf students (Scassellati et al., 2018).

When these pedagogical agents are used with students, regardless of if they play the role of tutors, coaches, or peers (Baylor and Kim, 2005), representation matters. Decisions for agents' appearance, language, and behavior may impact learners' perceptions of the cultural identity of the agents (Haake and Gulz, 2008), and may impact learners perceptions of their *own* belongingness and identity (*cf.* (Fordham and Ogbu, 1986)). Past work on agent representation also lacks alignment with modern understanding of identity, relying on binary definitions of gender (West and Zimmerman, 1987; Keyes, 2018) and failing to account for identities at the intersection of multiple marginalized groups (Crenshaw, 1990), especially in less developed countries (Wong-Villacres et al., 2018).

Incorporating representation improves embodied tutors, with improved student outcomes (Finkelstein et al., 2013). One of the simplest, most valuable steps for developers of ITS agents is to view the choice of the agent's identity presentation (identity factors such as race, appearance, voice, language, gender) as a non-neutral, political choice. The agents designed by researchers express to students beliefs about what a "model teacher" or "model student" look and sound like. Pracitioners and researchers alike often have great flexibility, at no additional expense, to intentionally design of the characters and content of the applications they create. This is different from the models themselves in a machine learning system, which rely on expensive training data, and which are often pretrained before development even begins, making it an attractive and high-leverage point for technologists to intervene.

5 Culturally Relevant Pedagogy

A lack of representation more broadly has contributed to an educational curriculum that privileges dominant cultures and which actively harms student engagement. The consequences are concrete - for instance, in recent bans on Chicano texts in the Southwest United States (Wanberg, 2013). One can draw a straight line back to historical policies that have devalued cultures, particularly for indigenous populations (Adams, 1995) and descendants of Black slaves (Alim et al., 2016; Lanehart, 1998). Historically, students coming from marginalized cultures have been measured by a "deficit model" (Brannon et al., 2008), where their home culture was viewed merely as a lack of knowledge about the dominant majority culture.

But there are alternatives in the existing pedagogy literature, like Moll et al. (2005)'s "funds of knowledge" model. This approach defines the accumulated and culturally developed bodies of information and skills that students learn at home and in their communities, essential to their functioning and well-being. An equitable approach

treats cultural knowledge instead as an asset, and allows students to build on what they know. This extends to technologies used in the everyday lives, homes, and communities of students - influencing their ability to impact student learning outcomes.

5.1 Case Study: Reading Comprehension

For early readers, speech recognition systems have been developed for children's voice and language (Gerosa et al., 2009) and are used to improve students' early reading skills (Mostow et al., 2003), or for speech-based vocabulary practice (Kumar et al., 2012). Yet these systems are often unable to generate questions for texts from nonstandard linguistic groups (*e.g.* with the syntactic and morphological transformations in African-American English (Siegel, 2001)). Systems today may also fail to recognize speech from students speaking certain dialects or accents, though progress in recognition for marginalized language variation is improving rapidly (Blodgett et al., 2016; Stewart, 2014; Jørgensen et al., 2015).

After basic literacy skills are acquired, NLP tools for language understanding are widely used to generate reading comprehension questions (Heilman and Smith, 2010). NLP is also used in related tasks like the measurement of readability (Aluisio et al., 2010; Vajjala and Meurers, 2012), and generation of simplified texts to differentiate homework based on student ability (Xu et al., 2015). But from a pedagogy perspective, content from these systems may be inappropriate - for instance, the questions generated are often factual rather than encouraging critical thinking (Rickford, 2001). This format does not measure student skills equally across cultures, and particularly under-reports progress in students of color, who tend to thrive when assessed through naturalistic narrative (Fagundes et al., 1998).

In pursuit of more reliable automated assessment, comprehension tasks may also fail to prioritize growth in student ability. Struggling readers understand texts more effectively when they are given chances to initiate dialogues and ask questions about texts, with teachers acting as listeners rather than ask their own questions about texts (Yopp, 1988). Teachers have difficulty creating these interactions (Allington, 2005), and intelligent agents have at least the potential for scaffolding tasks through real-time support for students as they perform their own tasks (Adamson et al., 2014). But to date, work has primarily focused on factoid assessment (Mostow and Jang, 2012; Zesch and Melamud, 2014; Wojatzki et al., 2016). This is an opportunity for future equitable NLP research at the intersection of ITS agents and reading comprehension. Additionally, coaching teachers to perform these dialogues has potential to fill in gaps in professional development and preservice training (Gerritsen et al., 2018), further incentivizing development of culturally responsive reading comprehension.

5.2 Case Study: Automated Writing Feedback and Scoring

Algorithmic assessment of student writing has taken many forms, from summative use in standardized testing (Shermis and Hamner, 2012) and the GRE (Chen et al., 2016) to formative use for classroom feedback (Woods et al., 2017; Wilson and Roscoe, 2019). This trend has led to sophisticated NLP analyses like argument mining (Nguyen and Litman, 2018) and rhetorical structure detection (Fiacco et al., 2019). Automated scoring has seen some more limited use in higher education, as well (Cotos, 2014; Johnson et al., 2017). For writers who are proficient or already working in professional settings, language technologies provide scaffolds like grammatical error detection and correction (Ng et al., 2014). These systems are enabled by rubrics, which give consistent and clear goals for writers (Reddy and Andrade, 2010). Rubric-based writing has drawbacks like rigid formulation of tasks (Warner, 2018), and many applications of rubrics are rooted in a racialized history difficult for technology to escape (Dixon-Román et al., 2019).

Bias creeps into rubric writing and scoring of training data, unless extensive countermeasures are taken to maintain reliability across student backgrounds and varied response types (Loukina et al., 2018; West-Smith et al., 2018). It also limits flexibility in task choice and response type from students, limiting students to writing styles that mirror the norms of the dominant school culture. Developers have an opportunity for equity work here, to the extent that they have leverage over task definition and training data collection (Lehr and Ohm, 2017; Holstein et al., 2018). Automated feedback systems may be improved through tasks that are flexible, and give culturally aligned opportunities for topic selection and choice; feedback

on rubrics that align to student "funds of knowledge" rather than the often-racialized language of deficits; and collaborative opportunities to share their work, receiving feedback that extends beyond algorithmic response.

6 Avoiding Linguistic Imperialism

Beyond selection of which content to teach, a broader issue is the focus of most language education globally on English and other prestige languages. This creates a privileged medium of communication and learning, and is rooted in colonialism; see for instance English's position over regional languages in India (Hornberger and Vaish, 2009) and the similar role of Afrikaans in South Africa (Heugh, 1995; Alim and Haupt, 2017); as well as how this extends to modern geopolitics in regions like Asia, with Han Chinese (He, 2013). In presumed-monolingual environments where students already speak the dominant language at home, this same effect plays out in dialects; examples include the privileging of white American or British dialects over stigmatized dialects like African-American Vernacular English in America (Henderson, 1996; Siegel, 2001), or the role of Classical Arabic as a prestige language over regional variants across the Arab world (Haeri, 2000). In language policy, this privileged position of a dominant language has been described as "linguistic imperialism" (Phillipson, 1992).

This dominant position of specific languages, especially English, comes despite cognitive science findings that bilingualism and code-switching ability has a marked positive effect on cognitive function (Petitto et al., 2012; Kroll and Bialystok, 2013) and may even have a positive economic effect on lifetime earnings (Agirdag, 2014). Moreover, language learning can promote new language acquisition while preserving respect for the learner's home language (or "heritage" language), helping learners to selectively choose when and how to communicate in each. Pedagogies exist which value pragmatic, socially conscious use of code-switching in mixed linguistic environments (Wang and Mansouri, 2017); these techniques are applicable to NLP.

6.1 Case Study: Computer-Assisted Language Learning

Computer-Assisted Language Learning, or CALL (Thomas et al., 2012), is an effective use of language technologies for vocabulary-building, pronunciation training, and practice through speech recognition, and other less common tasks (Witt, 2012; Levy and Stockwell, 2013). Language learning is a convenient fit for quantification, rapid experimentation (Presson et al., 2013), large dataset collection through "learner corpora" (Meurers, 2015), and fine-grained descriptions of progress through second language acquisition modeling (Settles et al., 2018). For second language teachers, NLP can improve their language awareness and skills (Burstein et al., 2014); for individual learners, language learning is highly personalizable and can be gamified for motivation and engagement (Munday, 2016). Machine learning models are also a good fit for summative assessment of student skill, and is used both in speech (Chen et al., 2018) and writing (Ghosh et al., 2016), including on high-stakes exams like the TOEFL (Chodorow and Burstein, 2004).

These systems make numerous design choices to implicitly or explicitly reject the grammar and lexicon of minority dialects. Typically, code-switching is neither taught as a skill nor supported as input. The relative sparsity of data for these variations may have resulted in unacceptable modeling accuracy in the past (Blodgett et al., 2016), but we are now closing that gap (Dalmia et al., 2018; Sitaram et al., 2019). For this field, an equitable language technologies agenda would seek to support rather than penalize these pragmatic skills. Such work can take place at multiple levels, beginning in early vocabulary work but particularly excelling in more sophisticated, scenario-driven practice for intermediate and advanced learners.

7 Surveillance Capitalism in Edtech

If we accept the premise that dominance hierarchies play a key role in education, it follows to acknowledge large-scale edtech that tracks students' activity in real time as one instantiation of "surveillance capitalism" in schools (Zuboff, 2015). Recent evaluations suggest that when students are aware of such systems in use, they report being anxious, paranoid, and afraid of long-term repercussions for undesirable behavior (Yujie, 2019). This may lead to short-term undesirable changes in students' behavior or expression to "game" algorithmic systems (*cf.* (Baker et al., 2008)). Effects may be greater in the long-term, with potential consequences to students' mental

health from always-on affect monitoring.

This presents an intersection for NLP to collaborate with information security and privacy researchers. Those fields are active in education, and the field has developed deep protections for students' personally identifiable data, enforced in America through laws like COPPA and FERPA (Regan and Jesse, 2018). While these laws do have gaps (Parks, 2017), they are largely robust and respected by technologists. More recent actions like the EU General Data Protection Regulation (GDPR) have also had meaningful impact on NLP research and data collection (Lewis et al., 2017). Legally, aggregating student data in order to develop and improve edtech provides a benefit to students and thus does not violate any law (Brinkman, 2013) — but scholars continue to ask ethical questions on how to account for student privacy and control (Morris and Stommel, 2018), and what data is being collected (Mieskes, 2017).

As always-on systems monitor students throughout their school day and beyond, these questions of student privacy and control become compounded in scope and complexity. Additionally, continuous monitoring impacts students' behavior and well-being: behavioral science has established that people change their actions when they are being observed (Harris and Lahey, 1982). Now, we must understand the impact when the observer is algorithmic.

7.1 Case Study: Student Engagement and Sentiment Analysis

One of the most common tasks in NLP research, for education and elsewhere, is sentiment and emotion recognition. This is important for education, both for design of affect-oriented curriculum (Taylor et al., 2017) and funding for socioemotional skills (Chan Zuckerberg Initiative, 2018). This recent turn is driven by promising initial results of efficacy from socioemotional interventions in schools (Dougherty and Sharkey, 2017). Measuring instantaneous student affective states is not only possible to reliably annotate, but also appears broadly possible to automatically infer (Yu et al., 2017); affect-aware tutoring systems are the subject of widespread research (Woolf et al., 2009; DMello and Graesser, 2012). In text-only settings online, sentiment has been a key part of prediction of attrition rates in MOOCs (Yang et al., 2013; Wen et al., 2014), especially when combined with micro-level instantaneous data like clickstream events (Crossley et al., 2016). These systems are now moving from data collected in text-only or tech-only environments, to multimodal data collected by always-on platforms like Alexa (Boccella, 2019) and emerging video monitoring platforms like the "Class Care System" (Yujie, 2019).

With this broad trend, we should question the implications of these systems as part of a move towards surveillance and monitoring, and their potential for impact on learners' well-being and behavior. Multimodal data are increasingly used to inform sentiment and affect detection algorithms (Yu et al., 2017; DMello and Graesser, 2012; Woolf et al., 2009), but these algorithms are known to produce discriminatory results, with disparate outcomes by gender (Volkova et al., 2013), race (Kiritchenko and Mohammad, 2018), and age (Díaz et al., 2018), perpetuating a quantifiable trend of disproportionate surveillance impact for people of color (Voigt et al., 2017). In a particularly illuminating example of bias introduced during corpus creation, Okur et al. (2018) found that experts from one culture radically misclassify affective states when they do not share the same cultural background as their subjects. A primary question for educational affect-detection systems will be to identify whether and how these discriminatory results replicate in educational systems, and will only become more urgent as real-time data from cameras, microphones, and other technologies become ubiquitous in the classroom.

8 An Equity Agenda

8.1 Representation on Teams

A theme of our review is that cultural representations should be built into NLP systems; here, though, we refer back to critical pedagogy's demand for cultural competence on the *builders* of these systems. Digital embodiment of characters from marginalized identities, developed by technologists without a background in those communities' culture and practices, runs significant risks of negative impacts and appropriation, or "digital Blackface" (Green, 2006). When NLP interventions mirror student cultures in purely performative ways, that representation is unlikely to be meaningful; indeed, it may worsen student engagement with agent-based systems. But these downfalls can be avoided through teams with "cultural competence" through lived experience and

group membership shared with the students they are building applications for.

A lack of diversity on research teams is a key contributor to discriminatory outcomes of machine learning systems in practice (West et al., 2019). Representational harms can be avoided by bringing those voices directly into the development of systems. Many of the challenges we have laid out are second nature to researchers with a cultural background in the communities that they seek to serve; having those voices in empowered positions during development can help make these issues salient before they are implemented - provided these voices are heard and empowered during the design process (*cf.* Holstein et al. (2018)).

8.2 Intentional Science Communication

As researchers, our work always has the potential to "go viral" and reshape public discourse. To illustrate, we can look to early language acquisition. In Hart and Risley (1995), researchers prominently reported findings of a "30-million word gap" for children raised in lower-class, predominantly Black households, hindering their literacy development. Later research showed this gap was likely overstated by an order of magnitude (Gilkerson et al., 2017), and likely excluded race-related environmental factors like bystander talk (Sperry et al., 2018). The discourse that emerged was largely discriminatory towards poor parents from minority backgrounds (Avineri et al., 2015).

But scientists can also cautiously understate results in public - most prominently in climate change policy and climate denialism (Dunlap, 2013). In other fields, collective action by researchers has produced unified stands on how their technology should be used ethically, as in the use of gene-editing tool CRISPR to modify unborn children - an action that evoked unified condemnation from governments (Collins, 2018), public figures (Lovell-Badge, 2019), and peer researchers in China[4]. Understanding the wider implications of research findings on NLP in education and positioning that work to have maximal impact is part of the job of effective science writing. Each circumstance is specific and there are no universal best practices - the key is to emphasize findings that are well-grounded in results, and to be intentional in how researchers encourage stories to evolve from those findings.

[4]https://www.yicai.com/news/100067069.html

8.3 Transparency and Regulation

If we do not take collective stances on ethical NLP in education from within our community, enforcement may instead come from external regulation. Some have argued this is a useful tool for enforcing accountability on algorithmic systems. Prior work has proposed regulatory frameworks that may serve as guidance (Whittaker et al., 2018); legal frameworks for these questions are already being developed (Kroll et al., 2016); bills are being introduced into the US Senate (Farivar, 2019). Potential outcomes include waiving trade secrecy for data science companies, or applying "truth-in-advertising" laws to AI systems. These may be general, or may prioritize specific focus areas like affect recognition.

Should we move in this direction, research will need to support regulation, improving transparency and governance of algorithmic predictions. NLP researchers have aggressively studied interpretability, offering explanation of results rather than predictions alone (Guidotti et al., 2018) - linguistic information is captured by newer neural language models of text (Conneau et al., 2018; Sommerauer and Fokkens, 2018) and speech (Elloumi et al., 2018; Krug and Stober, 2018), reading comprehension (Kaushik and Lipton, 2018), and machine translation (Shi et al., 2016; Raganato and Tiedemann, 2018; Belinkov and Glass, 2019). Other work focuses on replication, allowing consistent tying of modeling choices to changes in behavior (Dror et al., 2017, 2018).

But the connection to liability is rarely made explicit, and is worth emphasis. These tools are not just useful for error analysis and optimization of model performance; they will also be a critical step towards liability for harmful decisions made by algorithms, which cannot alter behavior if it cannot be traced and enforced (Ananny and Crawford, 2018). Governance can also come from somewhere in between collective action and national-level regulation. Some have proposed best practices for ethical industry research in NLP, mirroring IRB processes in universities (Leidner and Plachouras, 2017). This approach would assign responsibility during research, limiting experiments on users of commercial products. Either unregulated software will cause harm to students and teachers, or regulation and accountability to prevent inequitable use will come from somewhere. There is a spectrum of options for NLP,

from interpretability and self-governance to top-down regulation. It would be better for researchers to be at the forefront of that conversation.

8.4 Defining Boundaries for Software

As our last recommendation, researchers should acknowledge the "solutionism" trap endemic in technical research, which assumes that there is a methodological change that could fix any problem while maintaining the primacy of our algorithmic solutions (Selbst et al., 2019). Some activists advocate for leaving certain problems unresearched entirely, due to their intrinsic and systemic risk of harm for marginalized populations — see for instance this discussion in the case of facial recognition software, in Whittaker et al. (2018). Sometimes, machine learning systems will not be the right way to solve problems. A valuable contribution of future work will be to better lay out the taxonomies of ethics and equity that apply to NLP research, following work that has begun in algorithmic systems more broadly (Ananny, 2016). This will allow researchers to make consistent choices about which problems are tractable with technological solutions, rather than addressing each new problem in an ad hoc fashion (Chancellor et al., 2019). This can only improve the quality of the products we do choose to build.

9 Conclusion

Machine learning has made many promises that are going to be difficult to fulfill. Throughout the 1960s and 1970s, science fiction author Arthur C. Clarke described the aim of technology in education to be: *"Any teacher that can be replaced by a machine should be."* (Bayne, 2015). As late as 2015, adaptive learning companies like Knewton argued in favor of *"robot tutors in the sky that can semi-read your mind"* to replace traditional teachers (Westervelt, 2015). While this language has become more muted in recent years, the promise of AI and attached hype for our work is at an all-time peak. Language technologies in education have the potential to enable equity in the "pedagogical troika" of teaching, learning, and assessment (Gordon and Rajagopalan, 2016). While that potential is great, reifying existing power hierarchies is easy to do by accident or by choice; we hope researchers will resist simple answers, and build equity into future work from the start.

References

Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: quantifying stereotyping as a representational harm. *arXiv preprint arXiv:1901.09565*.

David Wallace Adams. 1995. *Education for Extinction: American Indians and the Boarding School Experience, 1875-1928.* ERIC.

David Adamson, Gregory Dyke, Hyeju Jang, and Carolyn Penstein Rosé. 2014. Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education*, 24(1):92–124.

Orhan Agirdag. 2014. The long-term effects of bilingualism on children of immigration: student bilingualism and future earnings. *International Journal of Bilingual Education and Bilingualism*, 17(4):449–464.

Vincent Aleven, Elizabeth A McLaughlin, R Amos Glenn, and Kenneth R Koedinger. 2016. Instruction based on adaptive learning technologies. *Handbook of research on learning and instruction. Routledge*.

H Samy Alim and Adam Haupt. 2017. reviving soul (s) with afrikaaps. *Culturally Sustaining Pedagogies: Teaching and Learning for Justice in a Changing World*, page 157.

H Samy Alim, John R Rickford, and Arnetha F Ball. 2016. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press.

Richard L Allington. 2005. *What Really Matters for Struggling Readers: Designing Research-Based Programs (What Really Matters Series)*. Boston, MA: Allyn & Bacon.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

Mike Ananny. 2016. Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, 41(1):93–117.

Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May*, 23.

Eleanor Armour-Thomas and Edmund W Gordon. 2013. Toward an understanding of assessment as a dynamic component of pedagogy. *The Gordon Commission, Princeton NJ*.

Netta Avineri, Eric Johnson, Shirley Brice-Heath, Teresa McCarty, Elinor Ochs, Tamar Kremer-Sadlik, Susan Blum, Ana Celia Zentella, Jonathan Rosa, Nelson Flores, et al. 2015. Invited forum: Bridging the language gap. *Journal of Linguistic Anthropology*, 25(1):66–86.

Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in gaming the system behavior in interactive learning environments. *Journal of Interactive Learning Research*, 19(2):185–224.

Ryan Shaun Baker and Paul Salvador Inventado. 2014. Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Anish Bavishi, Juan M Madera, and Michelle R Hebl. 2010. The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education*, 3(4):245.

Amy L Baylor and Yanghee Kim. 2005. Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(2):95–115.

Sian Bayne. 2015. Teacherbot: interventions in automated teaching. *Teaching in Higher Education*, 20(4):455–467.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2017. A dataset and classifier for recognizing social media english. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61.

Kathy Boccella. 2019. Alexa, can you help kids learn and teachers manage a classroom? at garnet valley, the answer is: Yes. https://www.philly.com/news/alexa-amazon-garnetvalley-school–20190213.html. Accessed 2019-04-13.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Lil Brannon, Jennifer Pooler Courtney, Cynthia P Urbanski, Shana V Woodward, Jeanie Marklin Reynolds, Anthony E Iannone, Karen D Haag, Karen Mach, Lacy Arnold Manship, and Mary Kendrick. 2008. Ej extra: The five-paragraph essay and the deficit model of education. *The English Journal*, 98(2):16–21.

Bo Brinkman. 2013. An analysis of student privacy rights in the use of plagiarism detection systems. *Science and engineering ethics*, 19(3):1255–1266.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

Jill Burstein, Jane Shore, John Sabatini, Brad Moulder, Jennifer Lentini, Kietha Biggers, and Steven Holtzman. 2014. From teacher professional development to the classroom: How nlp technology can enhance teachers' linguistic awareness to support curriculum development for english language learners. *Journal of Educational Computing Research*, 51(1):119–144.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Seth Chaiklin. 2003. The zone of proximal development in vygotskys analysis of learning and instruction. *Vygotskys educational theory in cultural context*, 1:39–64.

The Chan Zuckerberg Initiative. 2018. Chan zuckerberg initiative announces support to advance student development and success. https://chanzuckerberg.com/newsroom/chan-zuckerberg-initiative-announces-support-to-advance-student-development-and-success/. Accessed 2018-04-13.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (Atlanta GA*.

Jing Chen, James H Fife, Isaac I Bejar, and André A Rupp. 2016. Building e-rater® scoring models using machine learning methods. *ETS Research Report Series*, 2016(1):1–12.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based

automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238. IEEE.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Francis S. Collins. 2018. Statement on claim of first gene-edited babies by chinese researcher. https://www.nih.gov/about-nih/who-we-are/nih-director/statements/statement-claim-first-gene-edited-babies-chinese-researcher. Accessed 2019-04-13.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *Synthesis of tutorial presented at ICML 2018.*

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

Elena Cotos. 2014. *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. Springer.

Kate Crawford. 2017. The trouble with bias, 2017. *URL http://blog. revolutionanalytics. com/2017/12/the-trouble-with-bias-by-kate-crawford. html. Invited Talk by Kate Crawford at NIPS*.

Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241.

Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. 2016. Combining click-stream data with nlp tools to better understand mooc completion. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 6–14. ACM.

Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black. 2018. Sequence-based multilingual low resource speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4909–4913. IEEE.

John Dewey. 1923. *Democracy and education: An introduction to the philosophy of education*. Macmillan.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 412. ACM.

Ezekiel Dixon-Román, T. Philip Nichols, and Ama Nyame-Mensah. 2019. The racializing forces of/in ai educational technologies. *Learning, Media & Technology Special Issue on AI and Education: Critical Perspectives and Alternative Futures*.

Danielle Dougherty and Jill Sharkey. 2017. Reconnecting youth: Promoting emotional competence and social support to improve academic achievement. *Children and Youth Services Review*, 74:28–34.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Riley E Dunlap. 2013. Climate change skepticism and denial: An introduction. *American behavioral scientist*, 57(6):691–698.

Cynthia Dwork and Christina Ilvento. 2018. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Sidney DMello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157.

Sarah L Eddy, Sara E Brownell, Phonraphee Thummaphan, Ming-Chih Lan, and Mary Pat Wenderoth. 2015. Caution, student experience may vary: social identities impact a students experience in peer discussions. *CBELife Sciences Education*, 14(4):ar45.

Zied Elloumi, Laurent Besacier, Olivier Galibert, and Benjamin Lecouteux. 2018. Analyzing learned representations of a deep asr performance prediction model. In *Blackbox NLP Workshop and EMLP 2018*.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Deana D Fagundes, William O Haynes, Nancy J Haak, and Michael J Moran. 1998. Task variability effects on the language test performance of southern lower socioeconomic class african american and caucasian five-year-olds. *Language, Speech, and Hearing Services in Schools*, 29(3):148–157.

Cyrus Farivar. 2019. New bill aims to stamp out bias in algorithms used by companies. Accessed 2019-04-27.

James Fiacco, Elena Cotos, and Carolyn Rosé. 2019. Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 310–319. ACM.

Samantha Finkelstein, Evelyn Yarzebinski, Callie Vaughn, Amy Ogan, and Justine Cassell. 2013. The effects of culturally congruent educational technologies on student achievement. In *International Conference on Artificial Intelligence in Education*, pages 493–502. Springer.

Amber N Finn, Paul Schrodt, Paul L Witt, Nikki Elledge, Kodiane A Jernberg, and Lara M Larson. 2009. A meta-analytical review of teacher credibility and its associations with teacher behaviors and student outcomes. *Communication Education*, 58(4):516–537.

Signithia Fordham and John U Ogbu. 1986. Black students' school success: Coping with the burden of acting white. *The urban review*, 18(3):176–206.

Paulo Freire. 1970. *Pedagogy of the oppressed*. Bloomsbury publishing USA.

Paulo Freire. 1985. *The politics of education: Culture, power, and liberation*. Greenwood Publishing Group.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338. ACM.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Bill & Melinda Gates Foundation and the Chan Zuckerberg Initiative. 2019. Education research & development: Learning from the field.

Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. 2009. A review of asr technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, page 7. ACM.

David Gerritsen, John Zimmerman, and Amy Ogan. 2018. Towards a framework for smart classrooms that teach instructors to teach. In *International Conference of the Learning Sciences*, volume 3.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 549–554.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Edmund W Gordon and Kavitha Rajagopalan. 2016. Assessment for teaching and learning, not just accountability. In *The Testing and Learning Revolution*, pages 9–34. Springer.

Joshua Lumpkin Green. 2006. *Digital Blackface: The Repackaging of the Black Masculine Image*. Ph.D. thesis, Miami University.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93.

Magnus Haake and Agneta Gulz. 2008. Visual stereotypes and virtual pedagogical agents. *Journal of Educational Technology & Society*, 11(4):1–15.

Niloofar Haeri. 2000. Form and ideology: Arabic sociolinguistics and beyond. *Annual review of anthropology*, 29(1):61–87.

Jennifer Hammond and Pauline Gibbons. 2005. What is scaffolding? *Teachers voices*, 8:8–16.

Francis C Harris and Benjamin B Lahey. 1982. Subject reactivity in direct observational assessment: A review and critical analysis. *Clinical Psychology Review*, 2(4):523–538.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Baogang He. 2013. The power of chinese linguistic imperialism and its challenge to multicultural education.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617. Association for Computational Linguistics.

Lyn Henderson. 1996. Instructional design of interactive multimedia: A cultural critique. *Educational technology research and development*, 44(4):85–104.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129. ACM.

Kathleen Heugh. 1995. Disabling and enabling: Implications of language policy trends in south africa. *Language and social history: Studies in South African sociolinguistics*, pages 329–350.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239*.

John N Hooker and Tae Wan N Kim. 2018. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 130–136. ACM.

bell hooks. 2003. *Teaching community: A pedagogy of hope*, volume 36. Psychology Press.

Michael B Horn. 2018. Hey alexa, can you help kids learn more? *Education Next*, 18(2).

Nancy Hornberger and Viniti Vaish. 2009. Multilingual language policy and school linguistic practice: globalization and english-language teaching in india, singapore and south africa. *Compare*, 39(3):305–320.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

Adam C Johnson, Joshua Wilson, and Rod D Roscoe. 2017. College student perceptions of writing errors, text quality, and author characteristics. *Assessing Writing*, 34:72–87.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109. ACM.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):88.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *NAACL HLT 2018*, page 43.

Andrew J Kirkendall. 2010. *Paulo Freire and the cold war politics of literacy*. Univ of North Carolina Press.

Tanya Kraljic, Susan E Brennan, and Arthur G Samuel. 2008. Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107(1):54–81.

Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.*, 165:633.

Judith F Kroll and Ellen Bialystok. 2013. Understanding the consequences of bilingualism for language processing and cognition. *Journal of cognitive psychology*, 25(5):497–514.

Andreas Krug and Sebastian Stober. 2018. Introspection for convolutional automatic speech recognition. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 187–199, Brussels, Belgium. Association for Computational Linguistics.

Anuj Kumar, Pooja Reddy, Anuj Tewari, Rajat Agrawal, and Matthew Kam. 2012. Improving literacy in developing countries using speech recognition-supported games on mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1149–1158. ACM.

Rohit Kumar, Carolyn Penstein Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. *Frontiers in artificial intelligence and applications*, 158:383.

Gloria Ladson-Billings. 1995. Toward a theory of culturally relevant pedagogy. *American educational research journal*, 32(3):465–491.

Sonja L Lanehart. 1998. African american vernacular english and education: The dynamics of pedagogy, ideology, and identity. *Journal of English linguistics*, 26(2):122–136.

David Lehr and Paul Ohm. 2017. Playing with the data: What legal scholars should learn about machine learning. *UCDL Rev.*, 51:653.

Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by design: ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40.

Mike Levy and Glenn Stockwell. 2013. *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.

Dave Lewis, Joss Moorkens, and Kaniz Fatema. 2017. Integrating the management of personal data protection and open science with research ethics. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 60–65.

Diane Litman. 2016. Natural language processing for enhancing teaching and learning. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*.

Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov. 2018. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–12.

Robin Lovell-Badge. 2019. Crispr babies: a view from the centre of the storm. *Development*, 146(3):dev175778.

Nichola Lubold, Erin Walker, Heather Pon-Barry, and Amy Ogan. 2018. Automated pitch convergence improves learning in a social, teachable robot for middle school mathematics. In *International Conference on Artificial Intelligence in Education*, pages 282–296. Springer.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Detmar Meurers. 2015. Learner corpora and natural language processing. *The Cambridge handbook of learner corpus research*, pages 537–566.

Margot Mieskes. 2017. A quantitative study of data in the nlp community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29.

Robert J Mislevy, Pamela A Moss, and James P Gee. 2009. On qualitative and quantitative reasoning in validity. *Generalizing from educational research: Beyond qualitative and quantitative polarization*, pages 67–100.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Luis Moll, Cathy Amanti, Deborah Neff, and Norma Gonzalez. 2005. Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Funds of knowledge: Theorizing practices in households, communities, and classrooms*, pages 71–87.

Luis C Moll. 1992. *Vygotsky and education: Instructional implications and applications of sociohistorical psychology*. Cambridge University Press.

Ernest Morrell. 2015. *Critical literacy and urban youth: Pedagogies of access, dissent, and liberation*. Routledge.

Sean Michael Morris and Jesse Stommel. 2018. *An urgency of teachers: The work of critical digital pedagogy*. Hybrid Pedagogy Incorporated.

Jack Mostow, Greg Aist, Paul Burkhead, Albert Corbett, Andrew Cuneo, Susan Eitelman, Cathy Huang, Brian Junker, Mary Beth Sklar, and Brian Tobin. 2003. Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29(1):61–117.

Jack Mostow and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 136–146. Association for Computational Linguistics.

Pilar Munday. 2016. The case for using duolingo as part of the language classroom experience. *RIED: revista iberoamericana de educación a distancia*, 19(1):83–101.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

Behnaz Nojavanasghari, Charles E Hughes, and Louis-Philippe Morency. 2017. Exceptionally social: Design of an avatar-mediated interactive system for promoting social skills in children with autism. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1932–1939. ACM.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469.

Eda Okur, Sinem Aslan, Nese Alyuz, Asli Arslan Esme, and Ryan S Baker. 2018. Role of socio-cultural differences in labeling students affective states. In *International Conference on Artificial Intelligence in Education*, pages 367–380. Springer.

Django Paris and H Samy Alim. 2017. *Culturally sustaining pedagogies: Teaching and learning for justice in a changing world*. Teachers College Press.

Cecelia Parks. 2017. Beyond compliance: Students and ferpa in the age of big data. *Journal of Intellectual Freedom and Privacy*, 2(2):23.

Laura-Ann Petitto, Melody S Berens, Ioulia Kovelman, Matt H Dubins, K Jasinska, and M Shalinsky. 2012. The perceptual wedge hypothesis as the basis for bilingual babies phonetic processing advantage: New insights from fnirs brain imaging. *Brain and language*, 121(2):130–143.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web.

Robert Phillipson. 1992. Linguistic imperialism. *The Encyclopedia of Applied Linguistics*, pages 1–7.

Nora Presson, Colleen Davy, and Brian MacWhinney. 2013. Experimentalized call for adult second language learners. *Innovative research and practices in second language acquisition and bilingualism*, 38:139.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Y Malini Reddy and Heidi Andrade. 2010. A review of rubric use in higher education. *Assessment & evaluation in higher education*, 35(4):435–448.

Priscilla M Regan and Jolene Jesse. 2018. Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. *Ethics and Information Technology*, pages 1–13.

Angela Rickford. 2001. The effect of cultural congruence and higher order questioning on the reading enjoyment and comprehension of ethnic minority students. *Journal of education for students placed at risk*, 6(4):357–387.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of ACL*.

Noah Oluwafemi Samuel and Kate Williams. 2016. An english-language bibliography of the 1961 cuban literacy campaign. Technical report.

Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, et al. 2018. Teaching language to deaf infants with a robot and a virtual human. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 553. ACM.

Sylvia Scribner and Michael Cole. 1973. Cognitive consequences of formal and informal education. *Science*, 182(4112):553–559.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68. ACM.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65.

Mark D Shermis and Ben Hamner. 2012. Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual national council on measurement in education meeting*, pages 14–16.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

J Siegel. 2001. Pidgins, creoles, and minority dialect in education. *Concise Encyclopedia of Sociolinguistics, Elsevier, Oxford*, pages 747–749.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. https://arxiv.org/abs/1904.00784. ArXiv preprint. Accessed 2019-04-13.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Brussels, Belgium. Association for Computational Linguistics.

Douglas E Sperry, Linda L Sperry, and Peggy J Miller. 2018. Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child development*.

Ian Stewart. 2014. Now we stronger than ever: African-american english syntax in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37.

Holly E Tatum, Beth M Schwartz, Peggy A Schimmoeller, and Nicole Perry. 2013. Classroom participation and student-faculty interactions: does gender matter? *The Journal of Higher Education*, 84(6):745–768.

Rebecca D Taylor, Eva Oberle, Joseph A Durlak, and Roger P Weissberg. 2017. Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child development*, 88(4):1156–1171.

Michael Thomas, Hayo Reinders, and Mark Warschauer. 2012. *Contemporary computer-assisted language learning*. A&C Black.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173. Association for Computational Linguistics.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *LREC*.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

Tony Wan. 2019. Us edtech investments peak again with $1.45 billion raised in 2018. https://www.edsurge.com/news/2019-01-15-us-edtech-investments-peak-again-with-1-45-billion-raised-in-2018. Accessed 2019-04-13.

Kyle Wanberg. 2013. Pedagogy against the state: The ban on ethnic studies in arizona. *Journal of Pedagogy/Pedagogickỳ Casopis*, 4(1):15–35.

Hao Wang and Behzad Mansouri. 2017. Revisiting code-switching practice in tesol: A critical perspective. *The Asia-Pacific Education Researcher*, 26(6):407–415.

John Warner. 2018. *Why They Can't Write: Killing the Five-Paragraph Essay and Other Necessities*. JHU Press.

Miaomiao Wen, Diyi Yang, and Carolyn Rose. 2014. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational data mining 2014*. Citeseer.

Etienne Wenger. 2010. Communities of practice and social learning systems: the career of a concept. In *Social learning systems and communities of practice*, pages 179–198. Springer.

Candace West and Don H Zimmerman. 1987. Doing gender. *Gender & society*, 1(2):125–151.

Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems: Gender, race and power in ai. *AI Now Institute*.

Patti West-Smith, Stephanie Butler, and Elijah Mayfield. 2018. Trustworthy automated essay scoring without explicit construct validity. In *Proceedings of the AAAI Spring Symposium on AI and Society: Ethics, Safety and Trustworthiness in Intelligent Agents*.

Eric Westervelt. 2015. Meet the mind-reading robo tutor in the sky. http://tinyurl.com/y3l4jk4a. NPR Morning Edition. Accessed 2019-04-13.

John Wesley White. 2011. Resistance to classroom participation: Minority students, academic discourse, cultural conflicts, and issues of representation in whole class discussions. *Journal of Language, Identity & Education*, 10(4):250–265.

Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI now report 2018*. AI Now Institute at New York University.

Joshua Wilson and Rod D Roscoe. 2019. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, page 0735633119830764.

Langdon Winner. 1989. *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Silke M Witt. 2012. Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc. IS ADEPT*, 6.

Michael Wojatzki, Oren Melamud, and Torsten Zesch. 2016. Bundled gap filling: A new paradigm for unambiguous cloze exercises. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–181.

Marisol Wong-Villacres, Arkadeep Kumar, Aditya Vishwanath, Naveena Karusala, Betsy DiSalvo, and Neha Kumar. 2018. Designing for intersections. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, pages 45–58. ACM.

Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2071–2080. ACM.

Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. 2009. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14.

RE Yopp. 1988. Questioning and active comprehension. *Questioning Exchange*, 2(3):231–238.

Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1743–1751. ACM.

Xue Yujie. 2019. Camera above the classroom. https://www.sixthtone.com/news/1003759/camera-above-the-classroom. *Sixth Tone*. Accessed 2019-04-27.

Torsten Zesch and Oren Melamud. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Jan Henrik Ziegeldorf, Oscar Garcia Morchon, and Klaus Wehrle. 2014. Privacy in the internet of things: threats and challenges. *Security and Communication Networks*, 7(12):2728–2742.

Shoshana Zuboff. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89.