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 FATML2 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 technology3 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
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), 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 under-represented 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 (Goen 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 women’s 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. Practitioners 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 Mansour, 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 (Diaz 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 China4. 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.

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; Sommerrauer and Fokkens, 2018) and speech (Elloymi 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


Elena Cotos. 2014. Genre-based automated writing evaluation for l2 research writing: From design to evaluation and enhancement. Springer.


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.


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


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


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


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


