Defensible Explanations for Algorithmic Decisions about Writing in Education

Elijah Mayfield

CMU-LTI-20-012

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Alan W Black (Chair)	Carnegie Mellon University
Yulia Tsvetkov	Carnegie Mellon University
Alexandra Chouldechova	Carnegie Mellon University
Anita Williams Woolley	Carnegie Mellon University
Ezekiel Dixon-Román	University of Pennsylvania

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

Elijah Mayfield

Defensible Explanations for Algorithmic Decisions about Writing in Education

AUGUST 24, 2020

Carnegie Mellon University

PUBLISHED BY CARNEGIE MELLON UNIVERSITY

WWW.TREEFORTS.ORG

First printing, August 2020

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy.

Thesis Committee:

Alan W Black, Language Technologies Institute (Chair)
Yulia Tsvetkov, Language Technologies Institute
Alexandra Chouldechova, Heinz College of Public Policy
Anita Williams Woolley, Tepper School of Business
Ezekiel Dixon-Román, School of Social Policy & Practice
(University of Pennsylvania)

Abstract

This dissertation is a call for collaboration at the interdisciplinary intersection of natural language processing, explainable machine learning, philosophy of science, and education technology. If we want algorithm decision-making to be explainable, those decisions must be defensible by practitioners in a social context, rather than transparent about their technical and mathematical details. Moreover, I argue that a narrow view of explanation, specifically one focused on causal reasoning about deep neural networks, is unsuccessful even on its own terms. To that end, the rest of the thesis aims to build alternate, non-causal tools for explaining behavior of classification models.

My technical contributions study human judgments in two distinct domains. First, I study group decision-making, releasing a large-scale corpus of structured data from Wikipedia's deletion debates. I show how decisions can be predicted and debate outcomes explained based on social and discursive norms. Next, in automated essay scoring, I study a dataset of student writing, collected through an ongoing cross-institutional tool for academic advising and diagnostic for college readiness. Here, I explore the characteristics of essays that receive disparate scores, focusing on several topics including genre norms, fairness audits across race and gender, and investigative topic modeling. In both cases, I show how to evaluate and choose the most straightforward tools that effectively make predictions, advocating for classical approaches over deep neural methods when appropriate.

In my conclusion, I advance a new framework for building defensible explanations for trained models. Recognizing that explanations are constructed based on a scientific discourse, and that automated systems must be trustworthy for both developers and users, I develop success criteria for earning that trust. I conclude by connecting to critical theory, arguing that truly defensible algorithmic decision-making must not only be explainable, but must be held accountable for the power structures it enables and extends.

Contents

List of Figures

List of Tables

List of Abbreviations

For brevity in the sidenotes of this dissertation, my bibliography uses abbreviated acronyms for several of the most common publication venues in my field. In particular, you'll find the following shorthand:

Shorthand	**Full citation**
Proceedings of ACL	*Proceedings of the Annual Meeting of the Association for Computational Linguistics*
Proceedings of AERA	*Proceedings of the Annual Meeting of the American Educational Research Association*
Proceedings of AIED	*Proceedings of the International Conference on Artificial Intelligence in Education*
Proceedings of AIES	*Proceedings of the AAAI/ACM Conference on AI, Ethics,and Society*
Proceedings of BEA	*Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*
Proceedings of CHI	*Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*
Proceedings of CONLL	*Proceedings of the Conference on Computational Natural Language Learning*
Proceedings of CSCW	
(before 2017)	*Proceedings of the ACM Conference on Computer Supported Cooperative Work*
(2017-present)	*Proceedings of the ACM on Human-Computer Interaction: CSCW*
Proceedings of DIS	*Proceedings of the ACM Conference on Designing Interactive Systems*
Proceedings of EMNLP	*Proceedings of the Conference on Empirical Methods in Natural Language Processing*
Proceedings of FAccT	*Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*
Proceedings of Group	*Proceedings of the ACM International Conference on Supporting Group Work*
Proceedings of HICSS	*Proceedings of the Hawaii International Conference on SystemSciences*
Proceedings of ICLR	*Proceedings of the International Conference on Learning Representations*
Proceedings of ICWSM	*Proceedings of the International AAAI Conference on Web and Social Media*
Proceedings of LAK	*Proceedings of the International Conference on Learning Analytics and Knowledge*
Proceedings of NAACL	*Proceedings of the Annual Meeting of the North American Association for Computational Linguistics - Human Language Technologies*
Proceedings of NCME	*Proceedings of the National Conference on Measurement in Education*
Proceedings of NeurIPS	*Advances in Neural Information Processing Systems*
Proceedings of WikiSym	*Proceedings of the International Symposium on Wikis and Open Collaboration*

Table 1: List of publication venue abbreviations used in this work.

Part I: Goals

Introduction and Overview

"Any teacher that can be replaced by a machine should be."

— Arthur C Clarke[1]

"We think of it like a robot tutor in the sky that can semi-read your mind."

— Jose Ferreira, CEO, Knewton[2]

[1] Sian Bayne. "Teacherbot: interventions in automated teaching". In: *Teaching in Higher Education* 20.4 (2015), pp. 455–467

[2] Eric Westervelt. *Meet The Mind-Reading Robo Tutor In The Sky*. https://bit.ly/318Tj4b. NPR Morning Edition. Accessed 2020-08-01. 2015

The plan had been to write this thesis in coffee shops.

Or airports, or my office on campus. But instead, I wrote the last few articles for this dissertation while my coauthors and I were confined to our homes. For most of these last few months, the in-person economy of the United States has essentially been shut down. At the peak this spring, 165 countries had entirely closed their schools, with nearly 1.25 billion students impacted across primary, secondary, and tertiary education systems[3].

Although schools are closed, many are using technology to provide for continuity of learning. In higher education, many institutions shifted to distance learning quickly in the wake of campus closures, while K-12 systems have adapted in a variety of ways[4]. This is a critical time: Students will be impacted in enduring ways as the response of schools will likely exacerbate existing inequities. Staff will lose their jobs and institutional revenues will fall. Some campuses that have closed will never reopen.

What an opportunity for technologists. For a century or more, science fiction authors have imagined robots replacing teachers. Educational technology researchers and entrepreneurs have been eager to pick up this mantra and goal of replacing traditional teachers and schools. Over the last few years, the tone of these proclamations had started to die down. Yes, the promise of machine learning for enabling the "pedagogical troika" of teaching, learning, and assessment

[3] World Bank. *World Bank Education and COVID-19*. https://bit.ly/2yZwGFa. Accessed 2020-08-01.

[4] Justin Reich et al. "Remote Learning Guidance From State Education Agencies During the COVID-19 Pandemic: A First Look". In: *EdArXiv* (2020). https://doi.org/10.35542/osf.io/437e2

has held steady[5]. But airy buzz around superhuman predictive analytics in classrooms, textbooks, and school administrative offices had just started to fade into more reasoned discourse about capabilities and drawbacks, limitations, and implementation needs. That may now change. Some – perhaps most – school administrators will choose to confront this uncertain new reality by prioritizing the role of instructional technology in their students' learning. And my peers, the machine learning researchers that develop groundbreaking new algorithms for educational technology, will be eager to jump to respond to the call.

In response to Hurricane Katrina's impact on education in New Orleans, Naomi Klein wrote about how "disaster capitalists" such as Milton Friedman saw in that crisis an opportunity to "radically reform the educational system" through immense cuts to public education in favor of subsidies to for-profit private charter schools[6]. This spring, Klein wrote a reprise of this line of thinking - that she called the "Screen New Deal."[7] The coronavirus pandemic is a global crisis in nearly every sector, and the subsequent protests against police brutality are bringing anti-racist ideals to the forefront of our national discourse. The opportunity for rapid change in our social contract dwarfs even the furthest-reaching impacts of a localized natural disaster. This crisis is a once-in-a-generation moment to see new technologies adopted in sweeping fashion.

As I watch my colleagues in education and in technology, I see a disconnect between research and practice. In my career, I have worked with researchers, educators, funders, and policy-makers seeking to reshape schooling around educational technology. Along the way, it has been critical to defend the use of machine learning and natural language processing technologies. But the research on explainable machine learning in academic work today seems to struggle even on its own terms, much less survive contact with the outside world. This work answers few of the questions that users ask, while also missing the bigger picture of how tools are built in industry and used by downstream users (in education, this means teachers, students, staff, and parents). We are not well-equipped to explain algorithmic decision-making today, and I believe the reason is because questions about what makes a "good" automated decision are not just *technical*, they are *epistemological*.

[5] Edmund W Gordon and Kavitha Rajagopalan. "Assessment for Teaching and Learning, Not Just Accountability". In: *The Testing and Learning Revolution*. Springer, 2016, pp. 9–34

[6] Naomi Klein. *The shock doctrine: The rise of disaster capitalism*. Penguin Books, 2007

[7] Naomi Klein. "Screen New Deal". In: *The Intercept* (2020). Accessed 2020-08-01. https://bit.ly/3dZJhXw

Figure 1: Amid the 2020 coronavirus shutdown, New York's state government declared a plan to redesign their curriculum around online learning in collaboration with the Bill & Melinda Gates Foundation.

Problem Statement

How should developers and users understand algorithmic decision-making, making sense of automated choices and labels with stakes attached? There are options at every stage of implementation. Software developers will have to make detailed judgments of whether large-scale neural models are worth the cost and time investment over more straightforward baseline methods. In applied collaborations, domain experts will collect datasets; in doing so, they will have to decide on which training data to collect and how it should be labeled. Academics and corporations will try to defend the behavior of the models they train, and they'll need to be able to make clear and rational justifications. This requires good explanations of why the algorithms they trained are making trustworthy judgment calls. Here are my guiding concerns about these tasks:

1. **NLP research relies on narrow, deeply technical explanations of models and model architectures.** A rich, data-driven understanding of real-world behaviors informed by subject matter expertise is lacking in the literature, in favor of a search for causal explanation that I fear may be fruitless.

2. **We do not have a good bridge between machine learning research and software development.** Even builders of machine learning systems have few options for understanding the models we train; users have even fewer, and people impacted downstream by automated decisions often have no points of access at all.

3. **In education specifically, good intentions of machine learning researchers do not always translate to equitable impact.** Technical work is not steeped in the history of education reform, and as a result it is decontextualized from the consequences and fallout of automation.

I am not a skeptic of the entire field of educational technology. There's a long list of people who are, seeing the discipline as intrinsically harmful to schools[8,9,10]. I don't always agree with them – but leaning on algorithmic decision-making does require trusting the system from which that algorithm came. To build that trust, researchers must learn from their data, grappling with the human context of the domain they're working in. Rather than formal, causal guarantees, I argue for time-consuming work embedded in a broader social understanding of the actors that produced their data, understanding the culture you want to build, and training models that reflect behaviors you want to see replicated into the future.

[8] Audrey Watters et al. "The problem with'personalisation'". In: *Australian Educational Leader* 36.4 (2014), p. 55

[9] Ben Williamson. "Decoding ClassDojo: psycho-policy, social-emotional learning and persuasive educational technologies". In: *Learning, Media and Technology* 42.4 (2017), pp. 440–453

[10] Sharon Slade and Paul Prinsloo. "Learning analytics: Ethical issues and dilemmas". In: *American Behavioral Scientist* 57.10 (2013), pp. 1510–1529

Thesis Structure

Part I: Defining the Gap

I begin with an investigation of "explainability," as it is published in NLP. In partnership with philosophers of science, I dive into the specific scientific discourse of 2019, in the wake of the introduction of Transformers – particularly BERT[11]. We dive in specifically on the findings of Jain & Wallace[12] and Serrano & Smith[13], and the rebuttal by Wiegreffe & Pinter[14]. These arguments are nuanced and highly technical, and I try to take a step back and make a case from the humanities. I argue that attempts at interpreting deep neural models like BERT are *categorically* aiming for the wrong type of scientific explanation, in a way that is bound to get stuck in "traps" at the wrong level of target and abstraction[15]. The section ends with my reasoning for why we should lean instead on *non-causal* explanation of model behavior, and my goal for the thesis: to look at how we study and explain various applied domains that are relevant to education; in so doing, to begin looking for criteria for what defensible explanations should look like; and to tie these into a framework that can guide future NLP work.

[11] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL*. 2019

[12] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of NAACL*. 2019

[13] Sofia Serrano and Noah A Smith. "Is Attention Interpretable?" In: *Proceedings of ACL*. 2019

[14] Sarah Wiegreffe and Yuval Pinter. "Attention is not not explanation". In: *Proceedings of EMNLP*. 2019

[15] Andrew D Selbst et al. "Fairness and abstraction in sociotechnical systems". In: *Proceedings of FAccT*. ACM. 2019, pp. 59–68

Part II: Wikipedia Deletion Debates

As a first domain for building an approach to explainable decision-making, I spent time in 2019 understanding how predictions can be used to study the domain of real-world group debate. For this project, I did not start with an education problem that occurs directly in schools. Instead, I chose to study the rich discourse on deletion that rages behind the scenes of Wikipedia, the world's largest encyclopedia and source of open knowledge. I investigate the classification problem of predicting the outcome of text-based debates.

Female Nobel prize winner deemed not important enough for Wikipedia entry

Site moderator rejected submission for Donna Strickland, the first female physics winner in 55 years, in March

▲ Donna Strickland, the Nobel prize for physics winner at her home in Waterloo, Ontario, Canada, on 2 October. Photograph: Peter Power/Reuters

Figure 2: Contemporary news articles covered the deletion controversy around the recent Nobel laureate Donna Strickland. From *The Guardian*.

As a motivating example for why this domain is interesting, consider the popular, highly circulated story from the months I was coming back to CMU. At that time, news spread around the internet that Donna Strickland, an acclaimed scientist who had just won the Nobel Prize in Physics, had no Wikipedia presence; this was not only an omission but an *intentional decision*. She had, in fact, been subject of an article; but that article had been deleted due to falling below the standards of notability enforced by the site's editor community[16]. In this I found a highly inflammatory topic that comes up frequently for marginally famous individuals, and particularly people whose identity doesn't mirror that of Wikipedia's editor population. The site can be toxic to newcomers[17] and has an especially problematic and well-documented relationship with gender[18]. The impression was that the deck was stacked: Wikipedia's rules and culture tended to

[16] Leyland Cecco. "Female Nobel prize winner deemed not important enough for Wikipedia entry". In: *The Guardian* (2018). Accessed 2020-08-01. URL: `https://bit.ly/38YxvMt`

[17] Brian Keegan and Darren Gergle. "Egalitarians at the gate: One-sided gatekeeping practices in social media". In: *Proceedings CSCW*. 2010, pp. 131–134

[18] Shyong K Lam et al. "WP: clubhouse?: an exploration of Wikipedia's gender imbalance". In: *Proceedings of WikiSym*. ACM. 2011, pp. 1–10

shut out newcomers at the expense of longer-tenured members, and to prioritize the value of work from white, American men. But how do we describe and explain those social mechanisms?

I collected a massive corpus: every deletion debate in the history of Wikipedia since 2005, in total a little over four hundred thousand debates. My input is nearly-synchronous discussions where individual contributions are short but the discussion as a whole consists of contributions from multiple participants. I start by showing that this data can be used to train accurate machine learning models with BERT - we can do far better than guessing, and can build a useful machine learning tool for predicting the future. Given this ability to predict outcomes, I demonstrate that relatively simple models for classification tasks give us an avenue for reflection on the social phenomena that occur in the Wikipedia domain. I specifically show how trained models give us a rich quantitative set of tools for the human side of data science. The outputs of predictions give us a way to excavate insights from the group decision-making that *generated* the dataset. This process helps us to find out where decisions are actually coming from, and shows us what actions are predictive of success.

My specific investigation ends up focusing on the "calcified" rules and regulations that dominate debates over notability on Wikipedia[19]. These policies come from a limited set of a few dozen pages, mostly written over a decade ago, that drive debate among Wikipedia's editors[20]. Subjective biases and unfair actions, when encoded in algorithms, "become" objective, rule-based, almost robotic actions[21]. I'll use the predictions to explain how these policies are tied to group decisions, based on what outcomes the model forecasts.

While Wikipedia is not a school and does not have students or teachers, it intersects frequently with issues of curriculum design and learning. Open resources like Wikipedia undergird both the classroom and independent learning[22,23]. The discussions that editors have with each other about what belongs in those pages shapes knowledge itself – what gets to count as truth in their curriculum. These choices about content impact students directly. Students are perfectly aware of the cultural identity represented in the technology they use[24], and we have ample research going back decades that when students see (or do not see) themselves in the content of the books they read and the tests they take, it impacts their *own* sense of belongingness and identity in the school setting[25]. Furthermore, evidence continues to show that incorporating a better and broader representation accrues benefits for learning directly[26,27]. Given these factors, the impact of Wikipedia's editorial policy to teaching and learning is vitally important, and research on those policies has a clear connection on education more broadly.

[19] Brian Keegan and Casey Fiesler. "The Evolution and Consequences of Peer Producing Wikipedia's Rules". In: *Proceedings of ICWSM* (2017)

[20] Simon DeDeo. "Group minds and the case of Wikipedia". In: *Human Computation* (2014)

[21] Alexandra Chouldechova and Aaron Roth. "The Frontiers of Fairness in Machine Learning". In: *Workshop on Fair Representations and Fair Interactive Learning at the Computing Community Consortium* (2018)

[22] David Wiley and John Levi Hilton III. "Defining OER-enabled pedagogy". In: *International Review of Research in Open and Distributed Learning* 19.4 (2018)

[23] Matthew A Vetter, Zachary J McDowell, and Mahala Stewart. "From opportunities to outcomes: the Wikipedia-based writing assignment". In: *Computers and composition* 52 (2019), pp. 53–64

[24] Magnus Haake and Agneta Gulz. "Visual stereotypes and virtual pedagogical agents". In: *Journal of Educational Technology & Society* 11.4 (2008)

[25] Signithia Fordham and John U Ogbu. "Black students' school success: Coping with the "burden of 'acting white'"". In: *The urban review* 18.3 (1986), pp. 176–206

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

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

Part III: Automated Essay Scoring

Helping educators understand the role of machine learning in their work has been my job for a long time, with several different affiliations. The particular work I focused on was the commercialization of automated essay scoring (AES). This is a big industry: each year, millions of essays are scored automatically with models trained by machine learning, on exams like the GRE and GMAT[28]. Students write short essays of a few hundred words, usually on a specific writing activity with predefined content; an algorithm evaluates their work on a rubric based on past scoring data. My involvement in the field began in earnest in 2012 and 2013 with the publication of a Hewlett Foundation white paper[29], which circulated widely in the media and in policy-making circles. The evidence from that study established a belief, which has now been stable for several years: automated scoring is tractable with no more than standard machine learning methods[30].

[28] Yigal Attali and Jill Burstein. "Automated Essay Scoring with e-Rater® V. 2.0". In: *ETS Research Report Series* 2 (2004)

[29] Mark D Shermis and Ben Hamner. "Contrasting state-of-the-art automated scoring of essays: Analysis". In: *Proceedings of NCME*. 2012, pp. 14–16

[30] Mark D Shermis. "State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration". In: *Assessing Writing* 20 (2014), pp. 53–76

Figure 3: *The New York Times*' coverage of the edX EASE announcement drove much of the press attention to automated essay scoring in 2013.

But educators want more than scores on standardized tests. They want the reasoning behind those scores, and actionable advice for students to take away and use on future work. Maybe most importantly, they want to understand the role of the automated system in fundamentally human-human interactions like academic advising, teaching, and peer collaboration.

These extensions are not easy! The AES industry has historically

relied on multivariate regression models constructed by researchers with psychometrics expertise, defining and measuring only a few dozen intuitively satisfying and justifiable variables. The goal in building models has been not to attain the highest accuracy, but to weigh the tradeoffs between precision and recall, on one hand, against defensibility on the other. More modern NLP, including neural methods, are not amenable to this debate. Composition scholars who wish to explore AES are immediately overwhelmed with millions of parameters, hosted on GPU-enabled cloud compute, with no clear connection between the model's performance and any interpretable judgment system. The learning curve is simply insurmountable; today, practitioners must trust a black box from a hands-off developer.

So Part III of this thesis describes an approach to building the explainability case for an AES system. I start by questioning the importance of deep learning methods. The teams I led in industry used classical methods; they were based on ordinal logistic regressions and bag-of-words features, with minimal NLP technology; I describe the practical lessons learned from those years, including measurement of how well our software performed at improving student outcomes. In those same years, though, a massive influx of new tools became available to researchers, including neural networks, contextual word embeddings, and of course, attention-based Transformer architectures. It feels, intuitively, like those tools should be useful for essay scoring – but with serious technological requirements, a large carbon footprint, and no easy pathway to explainability, are they worth it? I explore that question, probing whether classical methods are good enough for AES tasks and whether full neural models are a necessary component for state-of-the-art research.

This section then moves on to a partnership with researchers implementing DAACS, the Diagnostic Assessment and Achievement of College Skills[31]. This support tool for first-time college students gives feedback on curriculum readiness for college as well as a wide range of self-regulation skills. Our goal was to build a good system for automated scoring, with behavior that could make sense to practitioners in live deployments at universities. I trained a series of classifiers for this product, in two waves of training data.

Thousands of students have used DAACS, but the population is unlike the target of most AES systems. Students are often mid-career, coming back to college in their 30s after a decade or more in the workforce, often in the military. From reviewing the text of their essays, I know that many have spouses, children, and jobs. Because of the potential for negative impact due to model bias[32], in addition to measuring performance of the models in terms of accuracy, I audit

[31] Diana Akhmedjanova et al. "Validity and Reliability of the DAACS Writing Assessment". In: *Proceedings of NCME*. 2019

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

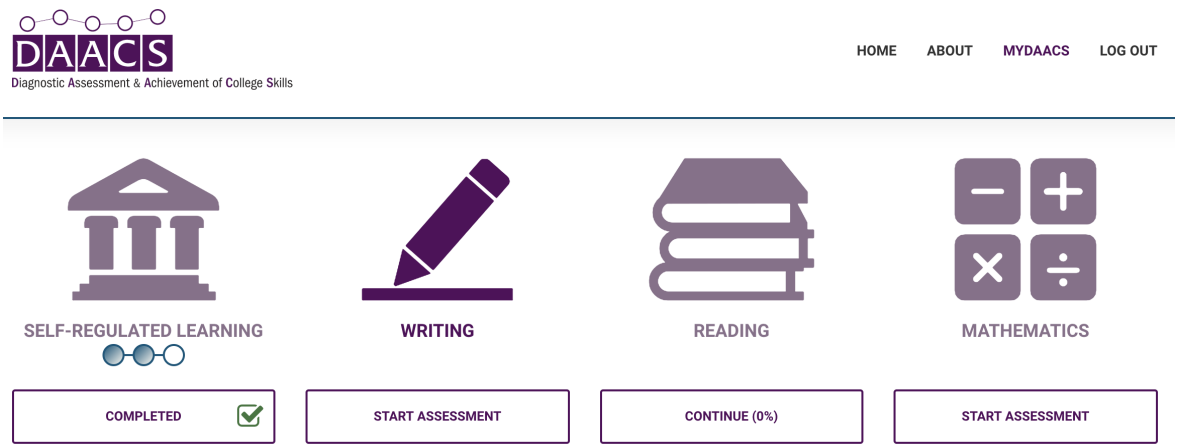


Figure 4: Homepage of the DAACS support tool for first-year college students.

the systems for demographic fairness by race and gender, using demographic data available through the DAACS partnership.

I also begin looking at the factors and features that might be part of a non-causal explanation for the scores that AES models produce. After all, an outstanding and highly controversial question in the essay scoring literature is exactly what patterns are prioritized by machine learning classifiers, and what implications that has for institutional adoption of technology, instructional pedagogy, and student test prep. Long-time critic of automation and former director of Writing Across the Curriculum at MIT, Les Perelman argues that algorithmic models learn trivial correlations with scores[33], like word count. Composition scholar Bill Condon argues a more subtle point, that the systems force students and institutions alike into valuing superficial writing styles, narrowing writing to an easily testable format[34].

So I take the methods used in the Wikipedia study and apply them to essay scoring. I try to explain the behavior observed in the live data from students using the system, in ways that might be pedagogically useful and productive for setting campus policy. I break this into two chapters: the first focusing on essay structure and in particular the use of the *five-paragraph essay*, and the second on essay content. I show how student writing can be a window into the social context of writing assessment and the normative contract between teachers, students, and institutions. Additionally, my content and topic analysis shows how personal expression, self-identity, and engagement with educational norms appear in the essays students submit to technological systems.

[33] Les Perelman. "When "the state of the art" is counting words". In: *Assessing Writing* 21 (2014), pp. 104–111

[34] William Condon. "Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?" In: *Assessing Writing* 18.1 (2013), pp. 100–108

Part IV: Toward Successful Explanation

With these two investigations, I spend the closing sections of the thesis defining success criteria for defensible machine learning decision-making. We can train models that reliably predict behavior, sure – but can we then explain the domain these models learn to emulate, the decision-making that our classifiers will bring into that domain, and the limits on the circumstances that our explanations will hold for? The final section of this thesis brings these questions together into a working conceptual frame for how future research should explain the human angle in machine learning data.

I lay out the case for non-causal explanations, again leaning on philosophy of science. I show how contemporary philosophers have developed the idea of *minimal models*, meant to define the shared features that, collectively, establish the circumstances under which a non-causal explanation will hold. I then describe how a *justificatory step* can carry the load for these explanations even when causality has not been established. And then, relying on work in computational social science, I describe how the data in both of my domains can be characterized. My justification for model explanations relies on the idea of individual humans as explainable social actors, with complex interactions and goals that algorithms learn to replicate. This gives a way forward for explainability research that eschews narrow, introspective explanations based on models alone in a vacuum.

But a condition of these explanations is that we know the systems that they can explain and the systems they cannot. In both of my domains of study, the social phenomena under analysis for algorithmic replication are based on systems of long-held structural biases; replicating and even explaining them may do more harm than good.

So I conclude by veering into more radical territory. The moment in 2020 is pivotal, and the stakes are incredibly high, but we have a problem. So far I've discussed how explainable machine learning researchers have focused on model internals rather than social context. Similarly, developers of education technology have focused on in-system analytics as the core of learning science[35]. But a natural extension of my work on social context is to critique the nature of how we measure algorithmic systems *in principle*. This final chapter, coming out of interdisciplinary reading on critical theory and queer feminist scholarship, further probes the goals and power structures of explainable machine learning, asking who benefits from automation in these domains, and what truths our explanations are actually revealing and reifying. I point out assumptions that are made when we build our systems in the first place. Now of all moments, we should question those assumptions.

[35] Ryan Shaun Baker and Paul Salvador Inventado. "Educational data mining and learning analytics". In: *Learning analytics*. Springer, 2014, pp. 61–75

Core Contributions

This dissertation is an opinionated call for *proactive* and *social* explainability in natural language processing systems. I argue for an explanatory strategy that can give usable insight and knowledge about algorithmic systems that is practical for real-world conversations about the consequences of automated decision-making.

- I make the case against a narrow view of explanation. I demonstrate that current approaches are not successful at interventionist causal explanation, but suggest hope that philosophy of science offers paths to successful non-causal accounts.
- I make contributions to two separate fields, using a broad toolbox for data-driven explanation of both human and model behavior.
  - **Wikipedia deletion debates**, where I present new data on group decision-making, release a large-scale corpus of structured conversations, and connect threads from prior work to gain insight about successful discourse strategies online.
  - **Automated essay scoring**, where I conduct an investigation of technical methods for automated essay scoring, measuring the tradeoffs of neural methods with classical approaches. Applying this approach to a dataset from first-year college students, I ask how we can describe and explain automated scoring behavior. My approach combines classification tasks with demographic fairness measurement and a mixed-methods investigation of both the structure and content of student writing.
- I propose a framework for developing defensible explanations for algorithmic decision-making, connecting the explainable machine learning literature to the social sciences and humanities. I also use this opportunity to connect to gender studies and recognize where an explainable algorithm must also be held accountable for the power structures it reifies and reinforces.

As researchers, we have a role to play in shaping the decisions that are being made about designing, developing, and disseminating automated decision-making systems. To responsibly fulfill that role requires a broader picture of how our systems are used and defended in practice. I set a course for knowing the behaviors of our models and the sources of those behaviors. Practitioners adopting these methods should be able to anticipate the consequences of adopting our technologies, not be surprised at first collision with the outside world. I hope to bring a new perspective on what it means to successfully build machine learning systems in education.

The Philosophy of Explanation

In the natural language processing community, we've reached a consensus that *explainability* in trained models is a positive attribute. When performing model selection, the less complex and more explainable model should be preferred (holding all else - for instance, classification accuracy or training time - equal). Part of this is purely intuitive and based on logistic ease for software developers; the preference for explainable models is also spurred on by regulation, led by the European Union's "right to explanation" in the 2016 enactment of the GDPR[36]. Yet so far, most researchers in NLP that describe their models as explainable have treated explanation the way that U.S Supreme Court Justice Potter Stewart famously treated obscenity: *"I know it when I see it"*[37]. It is challenging to evaluate the success of an explainable neural model without defining a criterion for evaluating what counts as an *explanation*, and moreover what counts as a *good* explanation. So let's start this thesis by asking what actually makes a good explanation for machine learning behavior.

[36] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine* 38.3 (2017), pp. 50–57

[37] Potter Stewart, *concurring*. "Jacobellis v Ohio". In: *United States Supreme Court* 378 (1964), p. 184

In language technologies, when publications aim to characterize or define explanation, authors tend to begin by assuming that there is some singular thing that is *an explanation*, and that other researchers will know it when they see it in the same way that the paper's authors do. Moreover, they assume that explainability can be measured against other measurable quantities in machine learning models, in the context of tradeoffs and holistic assessment of model quality. There is a long list of explanainable AI overview articles proposing desiderata, or desirable characteristics, of what such an explanation should look like[38]. But a central contention of other fields, most importantly philosophy of science, is that no such obvious phenomenon exists. This chapter represents a collaboration with philosophers of science that we published at LREC this year[39]. What I learned is that philosophers tend to look for boundaries or generalizable limits in science, more robust than what any one researcher or group might see from any one experiment or paper. When cross-referencing recent results in machine learning with accounts of scientific explanation from philosophy, we found that purely model-based quantitative explanations are unable to give causal explanations on their own.

[38] Zachary C Lipton. "The mythos of model interpretability". In: *ICML Workshop on Human Interpretabiltiy in Machine Learning* (2016)

[39] Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models". In: *Proceedings of LREC*. 2020

This is a problem for computer scientists! We typically believe that causal explanations are the pinnacle of good explanation. The mantra that *"correlation is not causation"* looms over our scientific

inquiry, description, and every discussion of machine learning model behaviors. In collaboration with philosophers at the University of Kentucky, I began to build a theoretical grounding of explanation, specifically by critiquing the explanations I tend to read at ACL. For that critique, I choose to use one particular theory of explanation: James Woodward's *manipulability through intervention*[40].

[40] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005

For philosophers, an *account* is an application of a philosophical theory to a scientific process, making explicit the set of assumptions and worldviews that are embedded in the actions, writing, or conclusions of the scientists under scrutiny. I begin this chapter with an overview of philosophical theories of scientific explanation, including brief summaries of multiple competing and complementary perspectives. This overview is itself a new contribution, both in the taxonomy of theories it develops and in its presentation for computer scientists in the explainable NLP community.

Next, from Woodward I'll specifically walk through the *interventionist account*, and evaluate the isolated type of explanations in the NLP literature, and present a philosophical foundation for a new and different style of explanation. Due to its emphasis on causal reasoning via counterfactual analysis and its historical development at the interface of computer science and philosophy of science, this account is a good tool for analyzing a class of recent findings on explanation of neural networks through causal reasoning.

Then I'll walk through some context on explanations in machine learning, stepping from rule-based systems to linear models and most recently to deep models. Then I'll dive in deep, describing one widespread and highly popular approach for explaining neural networks for tasks using text or speech inputs: the use of *attention mechanisms* as a functional basis for generating explanations. I then give a detailed summary of two recent findings on the limitations of attention mechanisms for explainable NLP.

- Jain & Wallace[41], which finds that attention layers in neural networks can be subject to adversarial reweightings, undermining their use for explanation.

[41] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of NAACL*. 2019

- Serrano & Smith[42], which finds that a very large number of attention weights can be zeroed out entirely, again undermining the use of these layers for identifying the importance of intermediate representations within a deep neural classifier.

[42] Sofia Serrano and Noah A Smith. "Is Attention Interpretable?" In: *Proceedings of ACL*. 2019

With an established background from both philosophy of science and deep learning, this chapter then gets to the meat of my critique. I apply the interventionist account to the study of attention mechanisms for explaining neural network behavior, focusing on three primary research questions:

RQ1: Is it appropriate to analyze deep learning explanations using the interventionist account?

RQ2: The account requires that interventions be surgical in order to make causal claims. Do the studies succeed at surgical intervention?

RQ3: If attention weights cannot be manipulated surgically, what are the consequences for explanation through attention?

To spoil the ending, I find narrow, model-only explanations lacking. But I end this chapter with some hope. The interventionist account deals only in *causal* explanation in the sciences. An important corollary of my analysis is that while a network will not and cannot produce causal explanations, it can still render alternate, non-causal types of explanations. I conclude the chapter by arguing that these types of explanations should be sought in the production of explainable neural models. I summarize these alternate types of explanation, making the key claim that non-causal explanations are the *only* types of explanation that can be derived from machine learning at the current state-of-the-art. I walk through the implications of these findings, pointing at multiple alternate accounts that suggest templates for success conditions for explainability with non-causal methods. Late in this thesis, after gathering evidence and examples from my two domains of interest, I'll use this background to suggest a path forward for explainability research.

Philosophy and Theories of Explanation

Philosophy of science research identifies and analyzes the conceptual and logical foundations of scientific reasoning using methodology including conceptual analysis, simulation modeling, formal methods, ethnography, and case studies on historical and contemporary instances of scientific research. The nature of scientific explanation has long been a topic of central concern in philosophy of science. Philosophical research on explanation seeks to identify what explanations are — whether they are instantiated patterns of logic inference, generators of a psychological sensation of understanding, ways of encoding similar patterns of information observed in disparate systems, traces of causes and effects, or something else entirely. Along with research on laws of nature, the structure of scientific theories, the aims of science, and the role of causation in the sciences, research on scientific explanation is one of the central subdisciplines within the philosophy of science.

The aim of philosophy is not consensus-building, so a wide variety of philosophical theories of explanation continue to coexist. Some are in direct competition with one another, while others serve as

	Theory	**Explananda** *(things to be explained)*	**Explanantia** *(things doing the explaining)*
Logical	Deductive-Nomological	Observed phenomenon or pattern of phenomena	Laws of nature, empirical observations, and deductive syllogistic pattern of reasoning
	Unification	Observed phenomenon or pattern of phenomena	Logical argument class
Causal	Transmission	Observed output of causal process	Observed or inferred trace of causal process
	Interventionist	Variables representing output of causal process	Variables representing input of causal process and invariant pattern of counterfactual dependence between variables
Functional	Pragmatic	Answers to why-questions	True propositions defined by their relevance relation to the explanandum they explain and the contrast class against which the demand for explanation is made
	Psychological	Observed phenomenon or pattern of phenomena	True propositions defined by their relation to the user's knowledge base and to the explanandum

Table 2: High-level overview of philosophical theories of explanation.

complements or limiting cases of one another. This brief review, summarized in Table 2, highlights a few of the most common sorts of theories of explanation, with emphasis on the varieties that are most central to explainability in neural models. A more comprehensive overview is available in Woodward's encyclopedia overview[43].

[43] James Woodward. "Scientific Explanation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University, 2017

Generally, theories of explanation may be understood as either *logical*, *causal*, or *functional*. Logical theories aim to characterize the logical structure of a cogent scientific explanation and typically emphasize the relations between explanation, laws of nature or scientific theory, and specific empirical observations. Causal theories aim to characterize explanation as an accounting of observed or expected patterns of cause and effect and are often accompanied by philosophical theories of causation itself. Functional theories, which typically focus on either the psychological or pragmatic functions of explanation, aim to characterize explanation in virtue of the function it accomplishes in scientific reasoning, rather than identifying the logical or causal structure of an explanation.

Some basic tenets of canonical theories of explanation are summarized below. Standard philosophical vocabulary for the parts of an explanation are employed: the *explanandum*, pl. *explananda*, is the thing to be explained, i.e. the target or object of an explanation; the *explanans*, pl. *explanantia*, is the thing doing the explaining.

- **Deductive-Nomological Theories**[44,45], one of the oldest logical theories of explanation, hold that explanations are deductive syllo-

[44] Carl G Hempel and Paul Oppenheim. "Studies in the Logic of Explanation". In: *Philosophy of science* 15.2 (1948), pp. 135–175

[45] Peter Railton. "A deductive-nomological model of probabilistic explanation". In: *Philosophy of Science* 45.2 (1978), pp. 206–226

gisms. The explanantia are the premises of the syllogism, and the explanandum is the conclusion. Among the explanantia, laws of nature are always taken as the major premise, and specific empirical conditions as the minor premise.

- **Unification Theories**[46,47], another logical theory, hold that explanations are not syllogistic; instead, they inhabit a more finely-structured logical space in which disparate phenomena exhibit the similar patterns of behavior. In this account, explanation consists in identifying the classification of a given argument pattern from among the accepted patterns of argument. The argument class is an explanans. Classes typically align with systems of natural laws.

[46] Michael Friedman. "Explanation and scientific understanding". In: *The Journal of Philosophy* 71.1 (1974), pp. 5–19

[47] Philip Kitcher. "Explanatory unification". In: *Philosophy of science* 48.4 (1981), pp. 507–531

- **Transmission Causal Theories**[48,49] characterize explanantia not as logical structures but as causal processes. These processes generate a product, which is the explanandum. Distinguishing genuine from merely apparent causal processes is a central concern of these theories and is accomplished by tracking the transmission of a signal, impulse, or mark over the course of the explanation.

[48] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984

[49] Phil Dowe. "An empiricist defence of the causal account of explanation". In: *International Studies in the Philosophy of Science* 6.2 (1992), pp. 123–128

- **Interventionist Theories** of *causation* introduce graph theory to the representation of causal relations and emphasize the identification of *invariance* relations between causes and effects as the target of causal claims[50,51,52]. Applied to *explanation*, the interventionist account[53,54] produces theories of causal explanation in which explananda and explanantia are connected by counterfactual causal dependence, which indicate invariant relations between purported causes and effects.

[50] James Woodward. "Capacities and invariance". In: *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grunbaum* (1994), p. 283

[51] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer-Verlag, 1983

[52] Judea Pearl. *Causality: models, reasoning and inference*. Springer, 2000

[53] James Woodward. "Explanation, invariance, and intervention". In: *Philosophy of Science* 64 (1997), S26–S41

[54] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005

- **The Pragmatic Theory**[55,56] contrasts itself with logical theories by characterizing explanation not as generation of a particular logical argument structure, and with transmission theories by not requiring the relay of a causal mark. Instead, the theory defines explanation functionally as answering a why-question about a phenomenon. Explanantia consist of meta-level logical structures that index explananda to an explanatory context and define relevance relations to contrasting phenomena.

[55] Bas C Van Fraassen. "The pragmatics of explanation". In: *American Philosophical Quarterly* 14.2 (1977), pp. 143–150

[56] Bas C Van Fraassen. *The scientific image*. Oxford University Press, 1980

- **Psychological Theories**[57,58] and their critics investigate explanation not as the satisfaction of any particular argumentative structure, but rather as acts or pieces of information that generate a sense of understanding in the agents (real or ideal) who interact with them. Significant attention is then given to characterizing what constitutes understanding. Like some of the work on explanation in neural models[59], this approach has significant overlap with the social sciences including psychology and behavioral science.

[57] Henk W De Regt. "The epistemic value of understanding". In: *Philosophy of Science* 76.5 (2009), pp. 585–597

[58] Kareem Khalifa. "The role of explanation in understanding". In: *The British Journal for the Philosophy of Science* 64.1 (2012), pp. 161–187

[59] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences". In: *Proceedings of ICJAI Workshop on Explainable AI*. 2017

The classification system here is meant to capture commonly-acknowledged divisions within philosophical research on explanation. Each of the theories identified above has benefits and drawbacks, and some are more appropriate than others for capturing the sort of explanation sought in the construction of explainable neural models.

My collaboration on the philosophical paradigms here was initially motivated by the observation that a significant source of confusion in explainable machine learning arises from failing to clarify what *sort* of explanation is being generated. For instance, an explanation for a model's output designed as causal fails if it only generates non-causal, nomological explanations. So we felt a need to define our terms and our conditions for success. When looking at the current state of most published machine learning research, we decided that the interventionist account was the best fit for what researchers in my field have been trying to do.

The Interventionist Account

This account focuses on those phenomena which can be explained in terms of the relationship between particular outcomes and the factors which gave rise to those outcomes. An explanation in this account relies on establishing the existence of *manipulability through intervention*. Some key features distinguish this account from others, such as logical explanations. First, the relationships between circumstances and outcomes are *empirical*, subject to data-driven verification through manipulation of those circumstances and collection of evidence. Next, that evidence is evaluated for *causality* - the dependence of the outcomes on additional variables is not merely conceptual but a direct relationship. The challenge here in explanation lies in concretely determining the existence of such a causal relationship.

In order to do so, the relationship between relevant variables in a system must be subjected to *manipulation*, where the values of those variables are changed. The theory thus lends itself well to explanations of systems which have quantifiable components, such that the quantity or value of the component can be easily denoted and modified as a variable. Performing a quantitative manipulation on a variable and then observing the changes in the output of the system as a whole, recording the overall changes through observed cause and effect, is called an *intervention*. If manipulation of quantifiable components leads to similarly quantifiable changes in output, researchers have established the first necessary, though not sufficient, elements of causal explanation.

The last piece of a successful causal explanation for a system re-

quires that an intervention on system components is *surgical*. To define this, philosophers lean on one final concept: *invariance*. In a multivariate system, it is frequently the case that a single effect has multiple causes. In Figure 5, the diagram on the left demonstrates a simple toy system with three variables: variable A is a causal factor for both B and C. B is also a causal factor in C. To perform a *surgical* intervention explaining the relationship between A and C, holding B invariant is necessary. But the system on the right, with only a handful of variables and relationships, demonstrates that some cases *resist* surgical intervention. The only path from D to H, for instance, is indirect, passing through other causal factors. We *cannot* hold those variables all invariant while intervening on D and still cause a change in H. According to Woodward, a *surgical* intervention is an intervention that makes strategic use of invariance; a successful explanation, finally, is only one that is generated empirically through the use of surgical interventions. The relationship between D and H cannot be explained through surgical intervention.

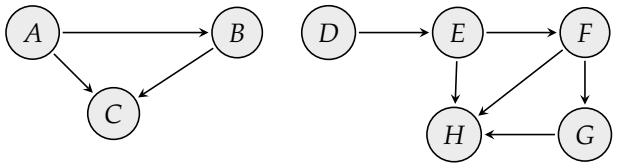


Figure 5: Network diagrams of causal systems. The system on the right resists surgical intervention between D and H.

The interventionist account is fundamentally *modal*: it relies upon counterfactuals in order to work:

> *"...an explanation ought to be such that it can be used to answer what I call a* what-if-things-had-been-different *question: the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways."*[60]

[60] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005

What the outcome of a manipulation would be, were it to occur, is what matters for a successful explanation: a pattern of counterfactual dependence between elements of the system of variables and the output of the system. In cases where we cannot track the pattern of counterfactual dependence among variables, no amount of manipulation is sufficient to successfully explain behavior. This has clear implications for neural models, which can have hundreds of millions of parameters, interdependent in complex ways.

The interventionist account is a good fit for probing the boundaries of causal explanation in machine learning, where inputs and outputs are quantifiable as vectors, tensor elements in neural models, or probability distributions in Bayesian models. Indeed, this definition of a successful causal explanation will be familiar to researchers with experience in Bayesian statistical machine learning.

Woodward's philosophical account was entwined with the development of Bayesian networks under his colleague Judea Pearl[61], and their shared research agenda led to mathematical definitions like *d*-separation of variables in machine learning. But while Pearl-style approaches to causality have been applied extensively in Bayesian models, their application remains daunting in the context of deep neural models.

Overall, the interventionist account has a great deal of appeal for computer scientists. Causal relationships are intuitive and aligns to how humans learn to interact with the world; if successful explanation requires human understanding on some level, there is great potential in an account that leverages natural inclinations to modify and test existing systems. But as we shall see, causality cannot always provide adequate explanatory power for large and complex systems.

[61] Dan Geiger, Thomas Verma, and Judea Pearl. "Identifying independence in Bayesian networks". In: *Networks* 20.5 (1990), pp. 507–534

Explainable Machine Learning

In machine learning, it is generally accepted that rule-based systems are easier to interpret by both amateurs and experts compared to linear models[62]; that linear models are in turn easier to interpret relative to generalized additive models[63] or Bayesian networks[64]; and so on until reaching the almost entirely black-box behavior of deep neural models[65]. There is room for translation, for instance by extracting simpler proxies that can mostly replicate more complex model behavior with simple rules[66]; but in general, explanation in machine learning has only gotten harder over the last more than forty years[67]. While this hierarchy is rudimentary and fails to account for all the various dimensions of model interpretability[68], it is broadly perceived to be accurate in practice. Nevertheless, a lack of explainability is rarely a barrier to implementation and widespread use. Neural models' performance continues to outpace other approaches to machine learning, and where they fall short in explanation, they make up for in performance.

[62] Himabindu Lakkaraju et al. "Interpretable & explorable approximations of black box models". In: *Proceedings of KDD Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2017)

[63] Yin Lou, Rich Caruana, and Johannes Gehrke. "Intelligible models for classification and regression". In: *Proceedings of the ACM SIGKDD*. ACM. 2012, pp. 150–158

[64] Carmen Lacave and Francisco J Díez. "A review of explanation methods for Bayesian networks". In: *The Knowledge Engineering Review* 17.2 (2002), pp. 107–127

[65] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* (2018)

[66] Longfei Han et al. "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes". In: *IEEE journal of biomedical and health informatics* 19.2 (2014), pp. 728–734

[67] Or Biran and Courtenay Cotton. "Explanation and justification in machine learning: A survey". In: *IJCAI-17 workshop on explainable AI (XAI)*. vol. 8. 2017, p. 1

[68] Zachary C Lipton. "The mythos of model interpretability". In: *ICML Workshop on Human Interpretabiltiy in Machine Learning* (2016)

But just because explanation of neural models has been difficult has not stopped researchers from trying. Much work has tried to grasp the structure of a network and what aspects of language are encoded where[69,70]. Others aim to measure how predictions change incrementally with new added information, evaluating the impact of each particular new input token in a text[71]. Additionally, in human-computer interaction researchers have worked to determine what users *want* from explanations[72], for instance by showing uses only a subset of text highlighted as important (as a simplifying step)[73]. This use of rationales, also known as *attributions*, can also be used directly at training time to encourage models to focus on or selectively ignore particular subsections of text[74,75]. This approach can be used without supervised span annotations. For instance, in my own prior work on essay scoring we observed how a model responded to word deletion[76] (and others have done similar things[77]). The results can be visualized using a variety of methods: heatmaps that perform highlighting or live editing[78]; generated plaintext, partially or entirely independent of the actual classification or factual content but facially plausible[79,80]; or direct exposure of structure in the underlying model, such as traversals through a graph[81].

[69] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "Bert rediscovers the classical nlp pipeline". In: *Proceedings of ACL*. 2019

[70] Kevin Clark et al. "What Does BERT Look At? An Analysis of BERT's Attention". In: *Workshop on Blackbox NLP at ACL*. 2019

[71] Jiwei Li et al. "Visualizing and Understanding Neural Models in NLP". in: *Proceedings of NAACL*. 2016, pp. 681–691

[72] Brian Y Lim and Anind K Dey. "Assessing demand for intelligibility in context-aware applications". In: *Proceedings of the International Conference on Ubiquitous Computing*. ACM. 2009, pp. 195–204

[73] Joost Bastings, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables". In: *Proceedings of ACL*. 2019

[74] Lucas Dixon et al. "Measuring and mitigating unintended bias in text classification". In: *Proceedings of AIES*. ACM. 2018, pp. 67–73

[75] Frederick Liu and Besim Avci. "Incorporating Priors with Feature Attribution on Text Classification". In: *Proceedings of ACL*. 2019

[76] Bronwyn Woods et al. "Formative essay feedback using predictive scoring models". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 2071–2080

[77] Dong Nguyen. "Comparing automatic and human evaluation of local explanations for text classification". In: *Proceedings of NAACL*. 2018, pp. 1069–1078

[78] Shusen Liu et al. "Visual interrogation of attention-based models for natural language inference and machine comprehension". In: *Proceedings of EMNLP*. 2018

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

[80] Hui Liu, Qingyu Yin, and William Yang Wang. "Towards Explainable NLP: A Generative Explanation Framework for Text Classification". In: *Proceedings of NAACL*. 2019

[81] Zhilin Yang et al. "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". In: *Proceedings of EMNLP*. 2018, pp. 2369–2380

Explanation through Attention

In recent years, much of the hope for explanation has been pinned on *attention mechanisms*. This innovation, first introduced by[82], allows neural models to be trained to automatically focus on small portions of inputs, like individual sentences or even words, while making predictions. This focusing allows neural models to outperform the state-of-the-art[83] and has led to sophisticated modern architectures like the Transformer, which has currently produced the most accurate models on a wide range of tasks[84,85]. In addition to performance gains, these layers appear to be providing human-interpretable explanations of model behavior "for free." Because the model was being trained to focus on specific subsets of information at inference time, the logic goes, it was reasonable to assume those dimensions are "most important" for the rationale of the resulting output. This approach resulted in a variety of visualizations and other attempts at model explanation being explored based, in part or in whole, on attention weights[86,87,88].

[82] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *Proceedings of ICLR*. 2015

[83] Diyi Yang et al. "Who Did What: Editor Role Identification in Wikipedia." In: *ICWSM*. 2016, pp. 446–455

[84] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of NeurIPS*. 2017, pp. 5998–6008

[85] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *Proceedings of ACL*. 2019

[86] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *Proceedings of the International Conference on Machine Learning*. 2015, pp. 2048–2057

[87] James Mullenbach et al. "Explainable prediction of medical codes from clinical text". In: *Proceedings of NAACL* (2018)

[88] Diyi Yang et al. "Let's Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms". In: *Proceedings of NAACL*. 2019, pp. 3620–3630

The past year has seen skepticism emerge about this indirect, downstream use of attention. The layer was designed to facilitate increased accuracy of models — in fact, the original paper makes no claims of its human interpretability — but the field saw wide proliferation of attention's use for explanatory purposes. In this section I briefly describe the two parallel studies, and one response paper, that serve as a foundation for our analysis.

"Attention is not Explanation" is an application of adversarial learning to the problem of explanation in machine learning systems. Jain & Wallace[89] take issue with the widespread direct extraction of attention weights into visualization tools like heatmaps. They show that *counterfactual* attention weights can be discovered for a given, trained neural network. First as a proof-of-concept, the authors show that attention weights can, in some cases, be randomly scrambled without loss of performance, suggesting that "explanations" derived from those weights have little meaning. They then demonstrate an optimization problem that moves attention weights *as far as possible* away from the original attention weights of a model, without changing the model's output behavior. The authors' critique of the use of attention for explanation describes these counterfactual configurations as equally plausible, from a modeling perspective, compared to other configurations which present far more intuitive explanations. Further, they assert a strong conclusion: because there exists the possibility that these adversarial configurations can be created without changing the outputs for given inputs, we *cannot* rely upon attention as a means of explanation.

[89] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of NAACL*. 2019

"Is Attention Interpretable?", written independently and contemporaneously, makes similar claims. Here, Serrano & Smith[90] test attention mechanisms by manipulating the layer's weights, and show that these weights do not impact output of the model. Rather than alter weights adversarially, the authors *omit nodes entirely*. They show that a surprising number of attention weights can be zeroed out (in some cases, more than 90% of nodes), without impacting performance of the model itself. The authors test this approach with a variety of ranking methods, from random choice to sophisticated sorting based on the gradient of nodes with respect to the classifier's decision boundary. Though their results show sensitivity to these different approaches, the core finding remains: Neural models are *highly* robust to change at the supposedly crucial attention layer, producing identical outputs in a large fraction of cases even after significant alterations to the model.

[90] Sofia Serrano and Noah A Smith. "Is Attention Interpretable?" In: *Proceedings of ACL*. 2019

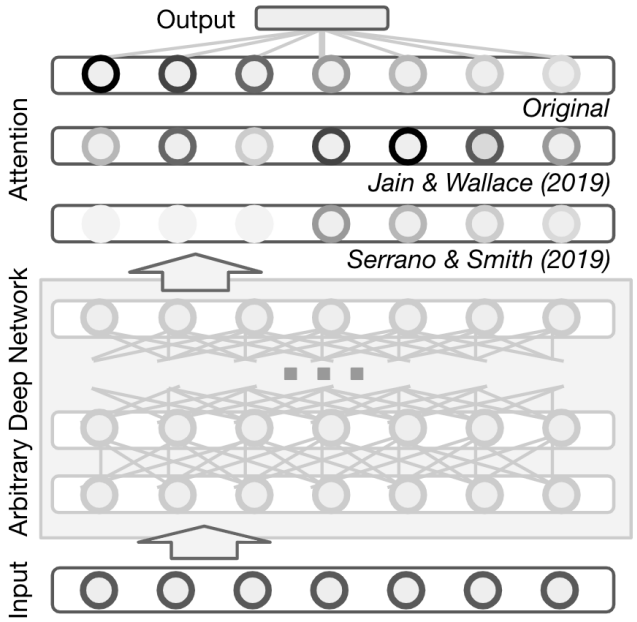


Figure 6: Researchers often use attention weights (top attention layer) to generate explanations. Jain & Wallace (middle) scramble weights and show that output remains stable; a similar result is obtained by Serrano & Smith (bottom) omitting highly-weighted nodes entirely.

But the picture is not simple. In *"Attention is not not Explanation"*, Wiegreffe & Pinter[91] produce several empirical results limiting the scope of the claims from the first two papers. The first result of that work shows that some of the classification tasks in the initial work are simply too easy for attention to matter — eliminating the layer by setting all values uniformly does not result in loss of performance. This suggests one practical boundary for attention by explanation, for any task where the additional complexity of an attention layer is wholly unnecessary to achieve state-of-the-art performance. For those tasks where attention is valuable for performance, the authors show

[91] Sarah Wiegreffe and Yuval Pinter. "Attention is not not explanation". In: *Proceedings of EMNLP*. 2019

that part of the vulnerability of manipulation to the attention layer comes from holding the rest of the model fixed. Attention as explanation, they argue, only makes sense in the context of a model that has jointly trained inner representation layers and the final attention layer. By constraining the adversary from Jain & Wallace to model-consistent behavior, they show that the resulting attention weights have much less room for modification without resulting in changes to the output.

As this debate goes on with additional empirical results, we find that computer science researchers are hampered by a lack of shared vocabulary and lack of a theoretical basis for success criteria of explanation[92]. So next, I advance the discussion by leaning explicitly on philosophy of science to build a more rigorous vocabulary and re-evaluate these results.

[92] Sam Corbett-Davies and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning". In: *Synthesis of tutorial presented at ICML.* (2018)

Applying the Interventionist Account

In what follows I will recast the findings from adversarial attention experiments from our highlighted studies in terms of Woodward's interventionist account, using the research questions introduced at the beginning of this chapter.

Does the Account Apply?

Woodward's interventionist framework may be a useful way to analyze attention mechanisms in NLP; but not every philosophical theory is an appropriate fit for every scientific experiment. Before proceeding I worked with my collaborators to confirm that the problem fits the conditions of a manipulation-based approach. In this case, we were looking for experiments that (1) produce empirical data, (2) hinge on causality as the core of their explanatory argument, and (3) rely on reasoning via counterfactual dependence to make causal claims.

In fact, adversarial attention configurations fit Woodward's description of intervention well. Woodward calls for modifying targeted variables in order to observe the changes to the output of the whole system, and adjustments to weights in the attention layer attempt just that: precisely shifting the focus of the algorithm to a particular segment of the input data in order to cause changes in the output, with the intent of measuring a causal effect (the outputs changing). When the system of variables represented by the attention configuration is manipulated, if there is a causal relation, the prediction generated by the model should change.

Jain & Wallace show that a vastly different attention layer which

results in the same output can be found by either searching through weights for nodes, or even by scrambling the weights of the network at random. The work from Serrano & Smith is similar. Here, rather than reweighting to create an adversarial layer, an enormous number of nodes in the attention layer can be zeroed out entirely, functionally removing them from the network. Again, they test whether outputs of the model differ based on this process. Both experiments evaluate the quantitative outputs of their models on a large corpus of data, meeting the requirement for empirical evidence as part of the explanation. Both also reason about counterfactuals to develop causal claims: the development of adversarial configurations of attention weights, or adversarially zeroed-out attention nodes, both are directly evaluated for explanatory value *compared to* the attention layer that was actually learned from data (the *base*, to use Wiegreffe & Pinter's terminology).

A1. Yes, these studies make arguments that fit the interventionist account.

Is Attention Manipulation Surgical?

Jain & Wallace's central proposition is clear from their title: attention is not explanation. They make a causal argument, discovering an adversarial attention configuration which produces the same effect, resulting in a loss of uniqueness of explanation. Furthermore, the resulting adversarial weights contradict intuitions about the sources of a model's judgment. Similarly, Serrano & Smith ask whether attention is interpretable. By showing that highly-ranked attention weights can be zeroed out without affecting model performance, they argue that the answer is no. These processes initially appear to be surgical intervention: researchers assert that the initial configuration is a plausible explanation, and after manipulation, the new configuration is implausible. Two major philosophical problems appear here.

First, it is possible that there actually is a true causal relationship in *both* the adversarial and non-adversarial attention configurations, and the model predictions. In this case, the original scientists would be right to conclude that attention is not explanation: a surprising, counterintuitive causal link between two variables that should not be linked is perplexing to users. In the colloquial sense, it explains nothing. On the other hand, an explanation of a causal relationship is not necessarily *non-explanatory* just because it is *unintuitive*. If manipulations reveal the existence of adversarial attention configurations that nevertheless produce accurate predictions, it may be the result of another causal relationship between inputs and the output class being predicted. If such a causal relationship does exist but was

not discovered through the model's original training, then yes, the counter-intuitive attention weights raise additional challenges and questions for researchers, who must then determine how and why these two variables are linked in this way. But this does not mean the adversarial explanation is *wrong*.

To get at the deeper problem, the interventionist account offers a second and more problematic observation on the experiments. In both original studies, attention is only one part of a larger system of variables; in fact, it is the final layer, receiving as input the result of a complex series of calculations on the initial inputs. But *invariance* in all non-target variables is what makes manipulations qualify as surgical. Both highlighted papers show an unsteady relationship between input tokens and the corresponding attention weights; the relevant variables for targets of manipulation lie outside of the attention layer. The relevant system in this case is not attention alone, but attention in addition to and in connection with the neural model's prior layers. If the generation of adversarial attention configurations is possible, the interventionist account argues, then there is more at work than attention in the learned model. This is quite a big problem to overcome, as the scope of the changes to network output of the network may not match the scope of attempted interventions. Additionally, engaging in interventions on selected weights does not result in continuous, smooth changes to model outputs, especially in discrete classification tasks.

Only some manipulations are surgical, and these example studies do not meet that standard; only some sets of variables are held invariant. Wiegreffe & Pinter make this argument implicitly in their response paper, arguing that severing the attention layer from the broader training of the overall model renders the experiment less meaningful; they argue that similar interventions on the remainder of the system are only possible with joint training between the attention layer and the rest of the model. Their critique is appropriate, and can be strengthened with philosophical vocabulary. Surgical interventions *require* explanations that depend on variables being held constant. Woodward's conditions for successful explanation are not met.

Failing this requirement has major consequences. The identification of causal relationships allows researchers to infer important details about the nature of the system which can serve in an explanation. But as the interventionist account cannot be meaningfully applied to systems which resist surgical interventions, causal explanations of the type that Woodward describes are not possible. Consequently, we will agree that attention is not explanation, but for reasons apart from, and broader than, the intuition-based arguments from Jain & Wallace. Instead we must argue that, *by definition*,

attention is not explanation. The manipulation of attention manipulations cannot meet preconditions laid out as part of the boundaries of successful causal explanation.

A2. No, manipulating attention weights fails conditions of surgical intervention.

Consequences of failed causal explanation

So alright. We cannot trace the causal chain through a neural model at the level of complexity in modern NLP. Woodward's framework demands the establishment of a pattern of counterfactual dependence through the elements of a system, and this can only be demonstrated through the use of surgical interventions while tracking changes in output. Without surgical intervention, we cannot determine if a pattern of counterfactual dependence exists in the first place, or if it does, how it is constituted. Two options present themselves:

- For some reason, Woodward's causal theory is inapplicable to attention-based manipulations, and the approach is not causally problematic.
- Woodward is correct, and the absence of the possibility for surgical intervention on attention mechanisms means that a causal link between attention and model output cannot be established.

In the first case, NLP researchers are faced with a difficult question: what distinguishes our circumstances from other quantitative systems of interacting variables in science, which are adequately explained by the interventionist account? But if we choose the latter, a more practical problem emerges: In order to be explainable, an algorithm must be manipulable via surgical intervention. The results of these papers suggest a failure in principle of modern NLP networks to allow for the testing of counterfactual manipulations. This categorically renders judgment that attention-based causal explanations are destined to fail on neural models. We do not have the ability to engage in surgical intervention on attention systems at all, and consequently cannot determine the nature of the causal relationships between attention layers and the output of the system. Though we cannot determine these relationships, we *can* still conclude that attention is not explanation — but only due to the broader claim that without access to and the ability to intentionally manipulate (or hold constant) all relevant system variables, explanation is ruled out entirely.

A3. Attention weights alone cannot be used as causal explanations for model behavior.

A constant in computer science, from calculating π with greater precision to mathematically complex but deterministic tasks like cryptography, has been that programs are constrained by the logic of their code, reliant on underlying notions of cause and effect. Deep learning cannot generate these causal explanations. But methods based in attention mechanisms *will* generate apparently causal explanations even where such reasoning is not possible.

Philosophical research has a grounding for these types of explanations: the *psychological* account. The success criteria for such explanations is not grounded in explanantia based in cause and effect, but in whether they produce a sense of understanding in the researcher or user of a system. The apparently causal nature of these explanations is in fact a hindrance to scientific understanding. In the context of manipulation where surgical intervention is not possible and psychological accounts take priority, the apparent causal stories are not reliable; they are causal fake news.

This undermines the central goals of explainable machine learning: to provide justification of why and how an algorithm made a decision, to hold the algorithm and its developers accountable for decisions that violate laws, and to give the subjects of those decisions actionable steps to alter the decision that the algorithm has made. If causal explanation is based in a failed methodology, all of the protections of explainability are suspect. Laws to protect members of marginalized classes will be enforced based on false understanding of model behavior; users seeking recompense will work in vain to alter their outcomes based on factors that will not produce change; and developers will allocate resources wastefully to improve model performance based on a misguided understanding of model behaviors.

But not all philosophical theories require *true* explanantia. While many theories do expect true explanantia and a testable, robust connection between explanantia and explananda, the door is opened for false but psychologically satisfying explanations. A strand of research in NLP has explicitly aimed not to generate *true* explanantia, but instead to produce *false* but *cognizable* explanantia: natural language generation, especially work using sequence-to-sequence modeling[93]. This direction of research would benefit from deeper vocabulary on the relation between truth and reasoning; as is, these explanations have no theoretical grounding of functional, logical, *or* causal theories. The challenge for such research will be to articulate the conditions for success of apparently causal explanations that are known to be false.

[93] Hui Liu, Qingyu Yin, and William Yang Wang. "Towards Explainable NLP: A Generative Explanation Framework for Text Classification". In: *Proceedings of NAACL*. 2019

Setting the Terms for Non-Causal Explanation

Philosophical Guidance

As models with interdependent relationships among a large number of variables grow, it becomes less likely that surgical intervention on variables can be performed. Moreover, even if such such surgical interventions were still possible in principle within the model, Wiegreffe & Pinter offer compelling concerns about the connection between attention layers and the broader model during training. To put it bluntly: the "deep" structure of contemporary NLP is exactly what prevents causal explanation from manipulation of their parts.

Nevertheless, the user affordances attached to many explanations employ causal vocabulary, despite such research being limited to purely psychological accounts of success. If meeting the success criteria for generating an explanation means producing human-cognizable systems of causal relations between variables, the point at which explanation becomes impossible is co-extensive with the point at which the number of variables in a causal chain exceed the maximum number of relations between variables which can, in principle, be tracked by a human to whom an explanation is directed. As the conclusion for our research, we argued that while researchers have defined their explanations informally using the constraints and success conditions of causal explanation, they are evaluating their success instead on non-causal theories of explanation, particularly either pragmatic or psychological bases.

The practical recommendation for NLP researchers is to disentangle explainability from cause-tracking. Rather than generating causally faulty (but psychologically satisfying) explanations that pattern themselves after causal explanation, as researchers have done in the past[94], we came to the conclusion that technical researchers of explanation should pattern their success conditions on *non-causal* accounts. Early in this chapter, I briefly identified non-causal accounts like the *logical* and *functional* types. These canonical philosophical theories of explanation should be part of explanation researchers' basic vocabularies.

[94] Upol Ehsan et al. "Rationalization: A neural machine translation approach to generating natural language explanations". In: *Proceedings of AIES*. ACM. 2018, pp. 81–87

But an even more promising body of non-causal explanation accounts come from contemporary philosophical research on explanation in mathematics and physics. These accounts of explanation are robust, under active study by philosophers, and are still available to NLP research:

- **Mathematical Explanations**[95,96], iterate on logical theories of explanation. Geometric explanations use mathematical principles as the explanantia, which are taken to be modally stronger than mere

[95] Christopher Pincock. "A role for mathematics in the physical sciences". In: *Noûs* 41.2 (2007), pp. 253–275

[96] Marc Lange. "What makes a scientific explanation distinctively mathematical?" In: *The British Journal for the Philosophy of Science* 64.3 (2013), pp. 485–511

causal principles or even natural laws of physics. A classic example of this type of explanation is the use of graph representations of the Bridges of Königsberg as the explanans for one's inability to cross all of the bridges exactly once in succession.

- **Structural Model Explanations**[97,98] identify scientific or mathematical models of systems as explanantia of the phenomena they represent. Explanations are built by connecting models to phenomena via a "justificatory step," whose details will be particular to the case at hand. This is a useful alternative framework for thinking about how explainable neural models will connect to the phenomena they aim to model. In early work, Sullivan[99] has begun evaluating the current prospects for deriving understanding from machine learning.

[97] Alisa Bokulich. "How scientific models can explain". In: *Synthese* 180.1 (2011), pp. 33–45

[98] Alisa Bokulich. "Searching for Non-causal Explanations in a Sea of Causes". In: *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations* (2018), p. 141

[99] Emily Sullivan. "Understanding from Machine Learning Models". In: *British Journal for the Philosophy of Science* (2019)

- **Minimal-Model Explanations**[100,101,102,103,104], drawing from work on the renormalization group in applied mathematics[105], generates a framework for justifying an explanation by using mathematical details to illuminate why differences between systems modeled via the same mathematics are irrelevant. By focusing on explaining away irrelevance, rather than articulating a relevance relation, these accounts flip the script for justification of purported explanations and produce a new theory of explanation of the functional sort. Due to its explicit engagement with explanations whose mathematics do not map cleanly onto represented features of the system being modeled, this approach may be especially promising for us in the context of text-based data and machine learning.

[100] Robert W Batterman. *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press, 2001

[101] Robert W Batterman and Collin C Rice. "Minimal model explanations". In: *Philosophy of Science* 81.3 (2014), pp. 349–376

[102] Collin Rice. "Moving beyond causes: Optimality models and scientific explanation". In: *Noûs* 49.3 (2015), pp. 589–615

[103] Collin Rice. "Models Don't Decompose That Way: A Holistic View of Idealized Models". In: *The British Journal for the Philosophy of Science* 70.1 (2017), pp. 179–208

[104] Collin Rice. "Idealized models, holistic distortions, and universality". In: *Synthese* 195.6 (2018), pp. 2795–2819

[105] Kenneth G Wilson. "Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture". In: *Physical review B* 4.9 (1971), p. 3174

When causal reasoning is taken off the table, some of the existing constraints from a causal conception are also placed at risk. Any non-interpretable neural network defies the sort of individuation of explanantia into a set of humanly-cognizable statements, premises, or causes, which is required for most of these theories. Because deep learning models defy this sort of individuation, recognizing which accounts of explanation work within deep learning will clarify evaluation criteria for explanation moving forward.

For the rest of this thesis, I aim to evaluate the success of a machine learning explanation on these justification-based accounts, rather than a causal account. The details of a good explanation are based on the justificatory step of domain expertise, followed by a structural model that covers the phenomena under examination. What does that look like in practice? The only way to know is to dig into the data.

Technical Innovations

Finding a new philosophical account by which to justify our explanations is one solution to the problems in this chapter, but it is not the only way out. An alternate approach is to reframe the technical problem being studied, and to seek out understanding and explanation of deep neural model behavior through alternate technical means. My work is hardly the only critique of attention-based explanation; adversarial actors can generated incoherent explanations with maliciously designed examples[106], sometimes requiring alterations to input text as small as a 1-character misspelling[107], and capable of actively deceiving users into believing explanations that are untrue or that omit key information in a purposefully deceitful manner[108].

As a result, researchers are pushing forward on alternate paths of explanation that rely less on standalone, interventionist analysis of attention weights, and instead use alternate paths toward what is now being referred to as *faithful* explanation. One common approach is detailed by joint work from the original authors of the articles analyzed in this chapter. In this work[109] and contemporaneous work by other researchers[110,111], researchers have sought to better define the goal of a *faithful* explanation. Stepping aside from a purely interventionist account, these works rely more on a psychological account, asking whether use of highlighted *rationales* in texts can lead to more faithful explanations based on the original intent of document labeling annotators. They suggest in their discussion that the next step for this branch of research, rather than formal proof of causality, is user study to evaluate whether the resulting explanations are sufficient for human understanding. In so doing they may make the case for moving the field toward a *trust*-based mindset; late in this thesis, I address this possibility and review what prior work from the human-computer interaction community may support such a turn.

An alternate approach to this problem would attempt to keep the causal rationale but move the unit of analysis away from individual tokens or neurons in a network. For instance, the approach of explanation through *influence functions*, first presented by Koh Liang[112], would alter the fundamental problem definition of explanation. Rather than focus on specific tokens or activation weights in a neural network, whether in attention layers or elsewhere, they argue for a focus on which *training examples* are most responsible for a given prediction. This premise was recently expanded by Han et al.[113] as a specific response to the theoretical gaps in an attention-based explanation strategy. In this work, they argue that not only are strictly attention-based models unsound for causal reasoning, they are nonsensical in more complex semantic tasks beyond document

[106] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of NAACL*. 2019, pp. 609–614

[107] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. "Learning The Difference That Makes A Difference With Counterfactually-Augmented Data". In: *Proceedings of ICLR*. 2020

[108] Danish Pruthi et al. "Learning to Deceive with Attention-Based Explanations". In: *Proceedings of ACL*. 2019

[109] Sarthak Jain et al. "Learning to faithfully rationalize by construction". In: *Proceedings of ACL*. 2020

[110] Julia Strout, Ye Zhang, and Raymond Mooney. "Do Human Rationales Improve Machine Explanations?" In: *Proceedings of the BlackboxNLP Workshop at ACL*. 2019, pp. 56–62

[111] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. "Fine-grained sentiment analysis with faithful attention". In: *arXiv preprint arXiv:1908.06870* (2019)

[112] Pang Wei Koh and Percy Liang. "Understanding Black-box Predictions via Influence Functions". In: *International Conference on Machine Learning*. 2017, pp. 1885–1894

[113] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. "Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions". In: *Proceedings of ACL*. 2020

labeling. These results show that not only is this effective at surfacing insights about training datasets, but that manually altered examples based on the training set produce predictable patterns of output that suggest the possibility of successful surgical intervention. This approach seeks to maintain the interventionist account while avoiding the pitfalls of attention-based explanation.

A final approach makes the smallest theoretical shift, arguing not that the interventionist approach is inherently flawed but that networks are merely overly dense in their relationships between nodes. Researchers such as Correia et al.[114] argue that explainable models may be achievable through intentionally induced sparsity, reducing the number of potentially divergent or malicious weightings that are available for spurious explanation. This approach focuses on practicality, culminating in models like DistilBERT[115], which is used later in this thesis.

[114] Gonçalo M Correia, Vlad Niculae, and André FT Martins. "Adaptively Sparse Transformers". In: *Proceedings of EMNLP*. 2019, pp. 2174–2184

[115] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019)

Ethical Limitations

Finally, a major, categorical limitation applies to all of the approaches I describe above. This definition of a successful explanation has all been about *epistemology*, not *ethics*. Many adjacent subfields of philosophy of science exist and only occasional interactions between computer scientists and philosophers have taken place to date[116,117]. In the chapters to come, I provide some foundation for how to build a good explanation. But a *good explanation* does not mean that a system has made a *good decision*; it certainly does not cover whether the system is using a *good algorithm*. Explainability research frequently studies tasks with high stakes, including notoriously biased tasks like recidivism prediction, financial risk modeling, and facial recognition for surveillance. Our work does not absolve researchers from a broader social responsibility: the presence of a successful explanation will not help if a loan is denied because of race[118], if an accused criminal is wrongly identified because of their gender presentation[119], or if algorithms persecute ethnic groups[120] or misdiagnose mental health[121].

Late in this dissertation, I argue that to build algorithmic decision-making in a truly socially responsible way, a successful non-causal explanation must be a component piece. But explainability, as a goal, must be incorporated into a much broader research agenda that accounts not only for explanation but also for ethical software development. This chapter is a good start: it provides a vocabulary for me to unify the informal language that proliferates across explainability research today, and allows the investigations that follow to maintain rigor even while avoiding making specifically causal arguments.

[116] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* (2018)

[117] John Zerilli et al. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" In: *Philosophy & Technology* (2018), pp. 1–23

[118] Andreas Fuster et al. "Predictably unequal? the effects of machine learning on credit markets". In: *The Effects of Machine Learning on Credit Markets (November 6, 2018)* (2018)

[119] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Proceedings of FAccT*. 2018, pp. 77–91

[120] Wei Wang, Feixiang He, and Qijun Zhao. "Facial ethnicity classification with deep convolutional neural networks". In: *Chinese Conference on Biometric Recognition*. Springer. 2016, pp. 176–185

[121] Cynthia L Bennett and Os Keyes. "What is the Point of Fairness? Disability, AI and The Complexity of Justice". In: *Workshop on AI Fairness for People with Disabilities at ACM SIGACCESS Conference on Computers and Accessibility*. 2019

Part II: Wikipedia Deletion Debates

Who gets to define what content is true and important? The answer to this question is often skipped in learning sciences, but this underlying argument over qualities of "good" knowledge and information eventually defines all the content that students end up seeing in the classroom.

In this first domain area for research, I dig into group decision-making as it occurs on the *Articles for Deletion* debates from Wikipedia's editor community. Far out of sight of most educators today, discursive processes between online users have crucial ramifications for students downstream as the curators of open knowledge. Central to a political, critical understanding of education is the idea that knowledge is not static, but is instead created through a discourse, and that the power dynamics that give structure to that discourse also directly impact the definition and creation of the knowledge that is produced. I show not only that those decisions can be predicted based on written debates, but that there's a better way of explaining the community's behaviors. I choose not to look introspectively at the features that are used by the model, trying to trace a causal path, but instead explain the human-to-human systemic behavior that generated the data on which those models are trained.

These sections of the dissertation were published in a pair of papers, first at the Computational Social Science workshop at ACL[122], and then in more extended form at CSCW[123]. The latter publication received a Best Paper Honourable Mention. I'll start with an overview of Wikipedia's context, the corpus that I built, and move on to the basic structure of the prediction task that I define. Then in the final chapter, I go on to show how we can explain community dynamics using those predictive models. This has both an immediate value for the Wikipedia research community, and creates a template for non-causal explanation in machine learning.

[122] Elijah Mayfield and Alan W Black. "Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making". In: *Workshop on Natural Language Processing + Computational Social Science at NAACL*. 2019

[123] Elijah Mayfield and Alan W Black. "Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–26

Context and Background

The decision of what counts as ideal and prestigious content, teaching, or learning is a high-leverage place for technologists to contribute. The rapid expansion of open educational resources shows signs of displacing the previous, publisher-driven model of textbook publishing[124]. These open practices appear to reduce gaps in student outcomes based on pre-existing income disparities[125] and students perceive the resources and curriculum as equal to, or better than, traditional textbooks[126]. Core to many of these resources is the use of material from Wikipedia, the free encyclopedia, assigning articles as readings and also actively writing content as part of course activities[127].

A Brief History of Wikipedia

Countless papers have studied Wikipedia (see Mesgari et al.[128] for a thorough survey), and a subset have studied editor interactions as a "model organism" for decision-making in online communities generally. This term, used in the context of social media analysis, originates with Tufekci[129]. Though Wikipedia was first founded in 2001, it took a few years for research interest in the community of editors to begin in earnest. The earliest published research on Wikipedia was likely[130]; shortly thereafter, numerous articles appeared in the following year. Of these, the comparison of Wikipedia's accuracy to *Encyclopaedia Brittanica* by Giles et al.[131] was most widely disseminated.

[124] David Wiley and John Levi Hilton III. "Defining OER-enabled pedagogy". In: *International Review of Research in Open and Distributed Learning* 19.4 (2018)

[125] Nicholas B Colvard, C Edward Watson, and Hyojin Park. "The Impact of Open Educational Resources on Various Student Success Metrics." In: *International Journal of Teaching and Learning in Higher Education* 30.2 (2018), pp. 262–276

[126] Rajiv S Jhangiani et al. "As Good or Better than Commercial Textbooks: Students' Perceptions and Outcomes from Using Open Digital and Open Print Textbooks." In: *Canadian Journal for the Scholarship of Teaching and Learning* 9.1 (2018), n1

[127] Robin DeRosa and Scott Robison. "From OER to open pedagogy: Harnessing the power of open". In: *Open: The philosophy and practices that are revolutionizing education and science. London: Ubiquity Press.* 4.1 (2017), p. 0

[128] Mostafa Mesgari et al. ""The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia". In: *Journal of the Association for Information Science and Technology* 66.2 (2015), pp. 219–245

[129] Zeynep Tufekci. "Big questions for social media big data: Representativeness, validity and other methodological pitfalls". In: *Proceedings of ICWSM*. 2014

[130] Andrew Lih. "Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource". In: *Proceedings of the International Symposium on Online Journalism, 2004*. 2004

[131] Jim Giles. *Internet encyclopaedias go head to head*. 2005

Much of this early work focused on editor motivation and hierarchy formation, trying to determine how high-quality writing could reliably emerge from spontaneous editor communities[132]; this early work built a foundational understanding of editor incentives and stratification that still informs much of today's research[133,134]. The growth in traffic caused a shift to maintenance work and internal debate rather than creation of new content[135]. This radical refocusing from content authoring to bureaucracy, led to a *"rational effort to organize"* through policies and guidelines[136]. While our work focuses on deletion debates, prior work has also studied deliberation and argument on in other administrative venues, like *Requests for Comment*[137] and on talk pages[138].

After this expansion period, the site experienced a long, steady decline in the following decade. The slowdown was noted almost immediately and attributed to three factors: an increase in overhead necessary for "maintenance" and administrative tasks for the larger community; newcomers turning away due to exclusion and gatekeeping from existing editors; and structural resistance to new edits through page protection and reverts[139]. The pattern of dropping activity continued for several more years, as the site matured and newcomer participation became even more difficult. While early decisions were made "by fiat" from user leaders or site founders, this was replaced over time by a decentralized network of editor committees, administrators, policies, and decision-making forums[140].

Today, much authority on the site remains grounded in a small network of policies shaped early in the site's history, written originally in response to a period of heavy growth and necessary crowd control. Some of the earliest policies, like Notability (`N`), `Verifiability` (`V`), and `No Original Research` (`NOR`, see Figure 7) originated many years ago but continue to dominate user revision activity and discussion, and drive group decision-making[141], while newer rules are comparatively obscure[142]. Wikipedia's norms for editor interaction are "highly conservative" and long-lived in comparison to most other online communities. These policies have "calcified," with newer policies falling to thrive and see broad adoption in later years; editors have instead favored iteration and refinement of a core set of pages[143]. This effect is partially due to newcomer behavior, which has trended away from spending time editing the site's core content in favor of time spent in discussion on talk pages, administrative disputes, and bureaucracy.

[132] William Emigh and Susan C Herring. "Collaborative authoring on the web: A genre analysis of online encyclopedias". In: *Proceedings of HICSS*. IEEE. 2005

[133] Andrea Forte and Amy Bruckman. "Why do people write for Wikipedia? Incentives to contribute to open–content publishing". In: *Proceedings of Group* (2005), pp. 6–9

[134] Fernanda B Viegas et al. "Talk before you type: Coordination in Wikipedia". In: *Proceedings of HICSS*. IEEE. 2007, p. 78

[135] Aniket Kittur et al. "He says, she says: conflict and coordination in Wikipedia". In: *Proceedings of CHI*. ACM. 2007, pp. 453–462

[136] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. "Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia". In: *Proceedings of CHI*. ACM. 2008, pp. 1101–1110

[137] Jane Im et al. "Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments". In: *Proceedings of CSCW* (2018), p. 74

[138] Khalid Al Khatib et al. "Modeling Deliberative Argumentation Strategies on Wikipedia". In: *Proceedings of ACL*. vol. 1. 2018, pp. 2545–2555

[139] Bongwon Suh et al. "The singularity is not near: slowing growth of Wikipedia". In: *Proceedings of WikiSym*. ACM. 2009, p. 8

[140] Andrea Forte, Vanesa Larco, and Amy Bruckman. "Decentralization in Wikipedia governance". In: *Journal of Management Information Systems* 26.1 (2009), pp. 49–72

[141] Simon DeDeo. "Group minds and the case of Wikipedia". In: *Human Computation* (2014)

[142] Bradi Heaberlin and Simon DeDeo. "The evolution of Wikipedia's norm network". In: *Future Internet* 8.2 (2016), p. 14

[143] Aaron Halfaker et al. "The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline". In: *American Behavioral Scientist* 57.5 (2013)

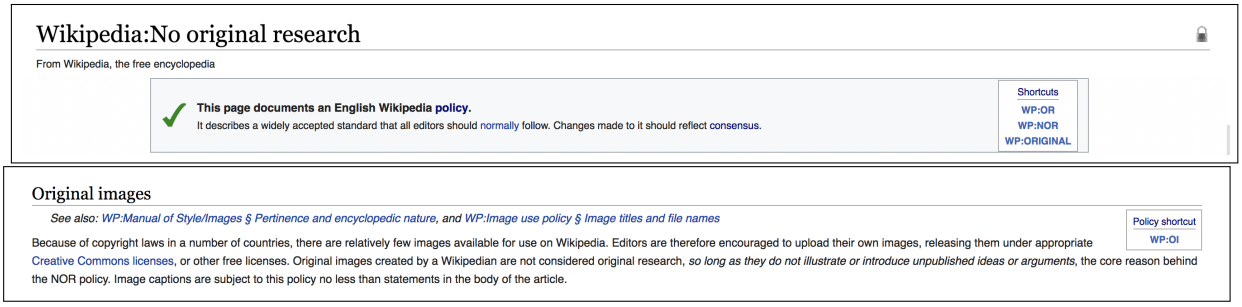


Figure 7: Top: Header of the `No original research` policy, which can be linked using aliases (`OR`, ,NOR, and `ORIGINAL`). Bottom: one specific subsection of that policy, which can be linked directly (`WP:OI`).

Articles for Deletion

In pursuit of better models for group decision-making, I chose to analyze Wikipedia's *Articles for Deletion* discussion domain. Editors at *AfD* nominate pages to these discussions when they believe they should be removed from the wiki, and usually include a nominating statement giving a rationale for deletion. After nomination, a discussion is held open for at least seven days. Exceptions to this timeline exist and allow "speedy" resolution of discussions - for instance, libelous pages or plagiarism of copyrighted material. When a page is nominated to *AfD*, any user (including unregistered users, provided they sign their post with an IP address) can place a vote, which must include a rationale for why they believe an article should be kept or removed from the wiki. These votes are public, signed, and timestamped. Users can also make non-voting comments, either in direct reply to the nomination, in reply to a vote or other comments. The structure of these comments follows the standard "reply tree" model of online discussion forums[144]. *AfD* is highly active, with more than one third of *all* articles in the English-language administrative namespace `Wikipedia:` related to deletion debates.

Discussions are aggregated by an administrator, who determines the discussion outcome. This is not a popular vote; the final tally of a debate is not the deciding factor, though administrators rarely deviate from consensus. Administrators may also hold debates open for a longer period of time, or close discussions with a verdict of `No consensus`. If no consensus is reached, nominated articles are kept by default; deleting articles requires an unambiguous outcome.

As with the rest of Wikipedia, *AfD* is subject to a broad set of written and unwritten norms for social behavior. Many of these norms have been encoded into hundreds of written and highly visible policies, guidelines, or essays. Note that these are terms of art, clearly denoted by page templates. Policies reflect broad, mandatory consensus, while guidelines contain generally accepted principles and essays are advice without broad acceptance (for more detail, see Forte

[144] Pablo Aragón et al. "Generative models of online discussion threads: state of the art and research challenges". In: *Journal of Internet Services and Applications* 8.1 (2017), p. 15

& Bruckman[145]). A long-running ideological divide in these debates exists on a spectrum between "deletionist" and "inclusionist." The former stance prefers high standards for material, culling less broadly relevant content and emulating the historical role of encyclopedias as gatekeepers. The latter stance argues for a reshaped role of information sources online, including, at its most extreme, *any* potentially valuable information that can be independently verified.

[145] Andrea Forte and Amy Bruckman. "Scaling consensus: Increasing decentralization in Wikipedia governance". In: *Proceedings of HICSS*. IEEE. 2008, pp. 157–157

The result was **keep**. Can't sleep, clown will eat me 01:21, 8 October 2007 (UTC)

Missed call [edit]

Missed call (edit | talk | history | links | watch | logs | views) – (View log)

Seems to fail WP:NOT, is essentially social commentary and no references are given for the major assertions presented. Orderinchaos 09:13, 2 October 2007 (UTC)

- **Delete** Just a junk article, not notable. Jmlk17 09:52, 2 October 2007 (UTC)
- **Keep**. I don't know guys, this thing is very prevalent in our culture. See [1]. I don't know for other cultures though. --**Lenticel** (talk) 10:18, 2 October 2007 (UTC)
  - Could possibly be mentioned in an article on Telecommunications in India? Twenty Years 15:50, 6 October 2007 (UTC)
    - Philippines is not India.--**Lenticel** (talk) 00:35, 8 October 2007 (UTC)
- **Keep** as above. I don't see why we should not keep this. .. Elmao 10:23, 2 October 2007 (UTC)
- **Keep**, there are bazillions of articles on cell phone etiquette out there to source this. I think the money-saving angle is only one part of it. --Dhartung | Talk 11:47, 2 October 2007 (UTC)
  - ...which could easily be covered in the article entitled Telecommunications in India. Twenty Years 15:57, 6 October 2007 (UTC)
- **Keep**, i added enough links to merit inclusion. it is not just a social commentary, it is a business, revenue and profit headache too. the apex body of indian telecom operators, coai has even instituted studies for tracking revenue loss. pls revisit the article to see the new links. Ankur Jain 12:23, 2 October 2007 (UTC)
  - **Comment:** hate to add this, but i believe there is a distinct anglo-american bias to article editing. just because you guys don't know about the widespread use of this thing, probably never having visited india or africa etc., that does not mean it does not exist. there is world beyond your countries.Ankur Jain 12:26, 2 October 2007 (UTC)

Figure 8: Excerpt from a single AfD discussion, with a nominating statement, five votes, and four comments displayed. Votes labeled in **"bold"** are explicit preferences (or stances), which are masked in our tasks.

Figure 8 gives an example of how these dynamics play out in practice for the article *"Missed Call."* The nominating statement cites the *"Wikipedia is not a dictionary"* policy and lack of sources to open the debate. This statement is followed by votes and comments, which also contain rationale texts. User preferences for `Delete` and `Keep` are given in **bold**, with some users voting to remove the page, some to keep, and discussion occurring through followup comments. After a long discussion and a total of eight votes and thirteen comments from ten total participants, the decision was made in favor of `Keep`.

Prior work on AfD debates

Substantial work on *AfD* has already taken place. The first detailed study of deletion decisions was conducted by Taraborelli & Ciampaglia[146]. This work found a herding effect among participants,

[146] Dario Taraborelli and Giovanni Luca Ciampaglia. "Beyond notability. Collective deliberation on content inclusion in Wikipedia". In: *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*. 2010, pp. 122–125

where later votes were highly influenced by the early tally of votes. It also found that user voting patterns could be well-described with a clustering model that contained only two clusters, and coarsely corresponded to "inclusionist" and "deletionist" users. The findings suggested significant biases in user behavior and made recommendations for more sophisticated analyses.

Next, a comprehensive early study attempted to directly quantify the *quality* of *AfD* debates[147]. They approached this problem by looking for articles that were deleted but later re-created, or kept but later re-nominated for deletion. They found a number of factors that led to good decision quality, like larger group size, groups that were diverse in experience level (but not groups heavy on recruited users or newcomers), and decisions made by unbiased administrators.

[147] Shyong K Lam, Jawed Karim, and John Riedl. "The effects of group composition on decision quality in a social production community". In: *Proceedings of Group*. ACM. 2010, pp. 55–64

Geiger and Ford[148] later analyzed debates and found a deep disconnect between the participants in debates and the authors that produced content. In particular, they found that an overwhelming majority of debates included no first-time participants at all, and that it was rare for article authors to participate in the discussion about their own article (under 20% of discussions). Later, Schneider et al. performed a qualitative review of 72 debates[149], conducting a close read of specific debates and recommending further research on the divide between readers and editors, obscure requirements and norms placed upon newcomer editors, and suggesting that the order of votes had a significant effect on debate outcomes.

[148] R Stuart Geiger and Heather Ford. "Participation in Wikipedia's article deletion processes". In: *Proceedings of WikiSym*. 2011, pp. 201–202

[149] Jodi Schneider, Alexandre Passant, and Stefan Decker. "Deletion discussions in Wikipedia: Decision factors and outcomes". In: *Proceedings of WikiSym*. ACM. 2012, p. 17

Following this work, Joyce et al.[150] tested a series of hypotheses on how rules and hierarchies interact with success in *AfD*. They replicated the prior finding that votes do predict outcomes, but added nuance on the use of seniority and policy, making several observations and doing a close study of two policy categories in particular, `Notability` and `Ignore All Rules`; the former was found to be universally predictive of successful votes, while the latter was correlated with success for `Keep` votes, but not `Delete`.

[150] Elisabeth Joyce, Jacqueline C Pike, and Brian S Butler. "Rules and roles vs. consensus: Self-governed deliberative mass collaboration bureaucracies". In: *American Behavioral Scientist* 57.5 (2013), pp. 576–594

More recently beginning in 2014, Xiao et al. undertook a series of mixed-methods studies of rationales in *AfD* votes[151]. They again replicated the finding that vote counts did significantly predict outcomes. Contradicting Joyce et al., they found that `Notability` topics were still the most common topic of argument in the domain but found no significant correlation between outcomes and the percentage of notability citations in discussions. They also surveyed *topics* for likely outcomes, and found significant relationships: biographies and for-profit companies were more likely to be deleted than other topics, while locations and events were more likely to be kept. Later work by the same researchers has avoided making outcome prediction a central research question, instead prioritizing discourse

[151] Lu Xiao and Nicole Askin. "What influences online deliberation? A Wikipedia study". In: *Journal of the Association for Information Science and Technology* 65.5 (2014), pp. 898–910

analysis: their most recent work has studied sentiment analysis[152], imperatives[153], and tree-style data visualization[154] for *AfD*, among other topics.

I situate my investigation in this body of work, with key results summarized in Table 3. To develop a more sophisticated model for analyzing the context of online debate, I begin my analysis with replication of specific findings from this prior work. I then move on to new observations. Our corpus contains a more comprehensive set of debates, both more recent and more thorough, nearly doubling the raw size of the largest prior studies. As a result, replication or failure to replicate may be a product of the larger sample size rather than a direct contradiction of past findings.

[152] Lu Xiao and Niraj Sitaula. "Sentiments in Wikipedia Articles for Deletion Discussions". In: *International Conference on Information*. Springer. 2018, pp. 81–86

[153] Lu Xiao and Jeffrey Nickerson. "Imperatives in Past Online Discussions: Another Helpful Source for Community Newcomers?" In: *Proceedings of HICSS*. 2019

[154] Ali Javanmardi and Lu Xiao. "What's in the Content of Wikipedia's Article for Deletion Discussions?" In: *Proceedings of The Web Conference (WWW)*. 2019, pp. 1215–1223

Prior Work	**Corpus**	**Key Findings**
Taraborelli & Ciampaglia (2010)	223k debates 2003-10	Early voters cause "herding." Individual users maintain `Delete/Keep` preferences across debates.
Lam et al. (2010)	158k debates 2005-09	Larger groups with a diversity of tenure produces better decisions. Recruiting creates biased groups but does not hurt decision quality. Bias of individual administrators can lower quality.
Geiger & Ford (2011)	120k debates 2007-11	Small groups dominate *AfD*. Article creators rarely participate. 96% of participation comes from repeat editors and 74% of debates have no newcomers.
Joyce et al. (2013)	588 debates pre-2012	Vote tallies and comment activity predict outcomes. Admin influence on outcomes is not significant. Citing the `WP:IAR` policy helps `Keep` votes.
Schneider et al. (2012-13)	72 debates, Jan. 2011	Novices and experts use different arguments. Both can be ineffective: novices make ineffective use of policy, while experts lean too much on boilerplate.
Xiao et al. (2014-19)	Subsets from 2010-2015 (229, 5k, 39k debates)	Notability dominates *AfD* rationales. Some topics, like biographies, have more unanimous outcomes than others. `Keep` votes have more positive sentiment. Expert editors frequently give imperative commands to newcomers.

Table 3: Summary of key findings from prior *AfD* studies. Our released corpus of 423k debates 2005-2018 contains a superset of all data in these papers, except early debates from 2003-04 in Taraborelli & Ciampaglia (2010).

Relevant Prior Work on Group Decision-Making

In group decision-making tasks, members participate in a constrained discussion, where they must choose from a fixed set of possible outcomes and there is no objective right answer[155]. Participants must debate the merits of the different choices and correctness is a judgment call, with persuasive arguments for multiple options[156]. In these tasks, dysfunction leads to poor outcomes, with low-quality discussion that fails to effectively fit together information from dif-

[155] Joseph Edward McGrath. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ, 1984

[156] Ignacio J Pérez et al. "On dynamic consensus processes in group decision making problems". In: *Information Sciences* 459 (2018), pp. 20–35

ferent group members[157]. High-performing groups by contrast have consistent characteristics like shared values, mental models, and communication styles, with nuanced patterns of conflict and consensus-building[158].

Group decision-making extends to online settings, where users in online production communities want to make good choices that will improve their collaboration over time. They accomplish this through intricate systems of social norms and cues for resolving conflicts[159]. But the details of how these decisions are made can be difficult to analyze or measure quantitatively. While much of online decision-making happens in free-form text discussions, much quantitative research ignores the details of this practice, *"observing change from before to after the deliberation without considering what has happened during the discussion"*[160]. Detailed analysis of discourse practices in the texts of online decision-making has seen less fruitful research activity compared to study of easier-to-quantify metadata or social network ties[161]. In behavioral science, questions are often explored through structured equation modeling and multivariate regressions, allowing behavior scientists sophisticated control over exogenous (fixed, external) variables, like demographics and task conditions[162], as well as *process* variables that describe observable behaviors in the groups being studied. Reducing team dynamics from text transcripts to quantitative process variables is computationally complex; in practice, text data is often ignored in favor of proxies like count statistics or, more frequently, participant survey responses[163].

A key reason is that getting at more sophisticated patterns is complex - most social science research instead avoids the question of extracting structure directly from text, instead relying on direct observable variables and survey data, or simulation[164], with explicit preferences encoded in modeled agents. In most work on group decision-making, automated discourse analysis is rare (with some notable exceptions, like work in some collaborative learning settings[165] and my own earlier work prior to working on automated essay scoring[166,167]). As a result, the proxies for explanation that come from much of the research on groups are reliable stand-ins, but put a limit on the *types* of questions that can be asked. Scientists studying teams may wish to evaluate which voices truly influenced a conversation, gauge the diversity of people or ideas represented in those influential roles, and measure observed conflicts and consensus-building. They may also want to assess whether any particular participant impacted the discussion and use these variables in aggregate to find which processes impact quality. This data is difficult to extract from discussion transcripts.

In the study of groups and teams, measuring discussion quality

[157] Wendy P Van Ginkel and Daan van Knippenberg. "Group information elaboration and group decision making: The role of shared task representations". In: *Organizational Behavior and Human Decision processes* 105.1 (2008), pp. 82–97

[158] Garold Stasser and William Titus. "Pooling of unshared information in group decision making: Biased information sampling during discussion." In: *Journal of personality and social psychology* 48.6 (1985), p. 1467

[159] Brian Keegan and Casey Fiesler. "The Evolution and Consequences of Peer Producing Wikipedia's Rules". In: *Proceedings of ICWSM* (2017)

[160] Jennifer Stromer-Galley and Peter Muhlberger. "Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy". In: *Political Communication* 26.2 (2009), pp. 173–192

[161] Dennis Friess and Christiane Eilders. "A systematic review of online deliberation research". In: *Policy & Internet* 7.3 (2015), pp. 319–339

[162] Gordon W Cheung and Rebecca S Lau. "Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models". In: *Organizational research methods* 11.2 (2008), pp. 296–325

[163] Daniel J Beal et al. "Cohesion and performance in groups: A meta-analytic clarification of construct relations." In: *Journal of applied psychology* 88.6 (2003), p. 989

[164] Francisco Chiclana et al. "A statistical comparative study of different similarity measures of consensus in group decision making". In: *Information Sciences* 221 (2013), pp. 110–123

[165] Carolyn Rosé et al. "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning". In: *International journal of computer-supported collaborative learning* 3.3 (2008), pp. 237–271

[166] Elijah Mayfield et al. "Computational representation of discourse practices across populations in task-based dialogue". In: *Proceedings of the International Conference on Intercultural Collaboration*. ACM. 2012, pp. 67–76

[167] Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. "Recognizing rare social phenomena in conversation: Empowerment detection in support group chatrooms". In: *Proceedings of ACL*. vol. 1. 2013, pp. 104–113

– plainly, what makes a group debate *good*? – is an open research area. Controlled behavioral studies have shown[168], for instance, that creativity, diversity, and conflict have major roles to play in the quality of teamwork. But the value of diverse discussion and open conflict is complicated, with a long history of positive, negative, and null results, depending on the narrow construct being studied[169]. What is clear is that the particulars of how teams are composed and how teammates interact with each other matters a great deal for effective group work[170,171].

Of course, large-scale corpus analysis is common in natural language processing, with many efficient representations of the complex underlying meaning of texts. Group discussion data is commonly used in NLP research. Datasets include the multiparty in-person group work of the AMI meeting corpus[172] and the pair task-based dialogues in the MapTask corpora[173]. In online contexts, group debates have been analyzed for tasks like argument mining[174] and stance classification[175], among others. Outside of NLP venues, though, most studies of groups and organizations do not perform sophisticated text mining or analysis. Methods vary: some research focuses on fuzzy logic or economic agent modeling[176], while others focus on social factors, network analysis, and the interactive aspects of teams[177,178].

Corpus Development

To explore this domain in detail, as part of the work for this proposal a large offline corpus of *Articles for Deletion* discussions. This snapshot contains the full text of all *AfD* debates in the English-language Wikipedia from January 1, 2005 to December 31, 2018. Prior to 2005, community norms, discussion formatting, and deletion process were more erratic, making automated extraction difficult and limiting any findings even if the data was successfully extracted (for this same reason, the corpus study in Lam et al.[179] also chose the January 2005 starting point). In addition to the raw text, this corpus is structured with extracted metadata, specifically timestamps, outcomes, nominations, votes, users, and policy citation. A total of 402,440 discussions were extracted. For the analyses that follow, I then filtered out two categories of discussions, mostly from earlier years in our corpus, when formatting norms were less standardized:

- Discussions without an outcome label from an administrator (20,669 instances, or 5.1%).
- Discussions that received no votes after nomination (12,179 instances, or 3.0%).

[168] Heather M Caruso and Anita Williams Woolley. "Harnessing the power of emergent interdependence to promote diverse team collaboration". In: *Diversity and groups*. Emerald Group Publishing Limited, 2008, pp. 245–266

[169] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. "Why differences make a difference: A field study of diversity, conflict and performance in workgroups". In: *Administrative science quarterly* 44.4 (1999), pp. 741–763

[170] Frances J Milliken, Caroline A Bartel, and Terri R Kurtzberg. "Diversity and creativity in work groups". In: *Group creativity: Innovation through collaboration* (2003), pp. 32–62

[171] Steve WJ Kozlowski and Daniel R Ilgen. "Enhancing the effectiveness of work groups and teams". In: *Psychological science in the public interest* 7.3 (2006), pp. 77–124

[172] Iain McCowan et al. "The AMI meeting corpus". In: *Proceedings of the Conference on Methods and Techniques in Behavioral Research*. Vol. 88. 2005, p. 100

[173] Anne H Anderson et al. "The HCRC map task corpus". In: *Language and speech* 34.4 (1991), pp. 351–366

[174] Fiona Mao, Robert Mercer, and Lu Xiao. "Extracting imperatives from wikipedia article for deletion discussions". In: *Proceedings of the Workshop on Argumentation Mining at ACL*. 2014, pp. 106–107

[175] Parinaz Sobhani, Diana Inkpen, and Stan Matwin. "From argumentation mining to stance classification". In: *Proceedings of the Workshop on Argumentation Mining at NAACL*. 2015, pp. 67–77

[176] Ignacio J Pérez et al. "On dynamic consensus processes in group decision making problems". In: *Information Sciences* 459 (2018), pp. 20–35

[177] John M Levine, Lauren B Resnick, and E Tory Higgins. "Social foundations of cognition". In: *Annual review of psychology* 44.1 (1993), pp. 585–612

[178] J Richard Hackman. *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers, 2011

[179] Shyong K Lam, Jawed Karim, and John Riedl. "The effects of group composition on decision quality in a social production community". In: *Proceedings of Group*. ACM. 2010, pp. 55–64

	Delete	Keep	Merge	Redirect	Other
Votes (2005-2018)	54.9	28.4	3.6	3.8	9.3
Outcomes (2005-2018)	63.9	20.7	3.2	6.0	6.2
Prior Work (Taraborelli & Ciampaglia) (2003-2010)	63.6	23.6	3.9	1.9	7.0

Table 4: Overall breakdowns of labels across all data.

After these exclusions, our analysis covers 369,592 debates. Table 4 shows percentages for each label for vote and outcome distributions in the analyzed subset. To analyze policy norms, I manually assembled a list of frequently cited links in *AfD* discussions. Editors can link to overall policy pages or directly to subsections; additionally, many pages and subpages can be linked using any of a number of shortcut aliases. The taxonomy I built includes 37 policy pages with 377 sections, 44 guideline pages with 398 sections, and 71 essay pages with 201 sections, all linked by a total 2,111 aliases. For each contribution, I extract all hyperlinks to any one of the aliases in our taxonomy. While this is a reasonable proxy, there are three reasons why this is not comprehensive:

- Most citations are added intentionally by the editor who signs the contribution; however, some are added after the fact (like links to the `SIGNATURES` policy, appended by bots to unsigned posts, along with the username or IP address logged for the contribution).
- While this taxonomy includes all official policies and the vast majority of guideline and essay citations in *AfD*, there is a long tail of rarely-cited essays and pages that are not comprehensively included in our taxonomy and were not extracted.
- I do not capture citations to policies without MediaWiki links to those pages (merely writing "NBIO" to refer to the notability policy on biographies, for instance, instead of writing "[[WP:NBIO]]" to include a link). Editors generally follow formatting conventions and include links when appropriate, but this is a source of missing data.

Corpus Preprocessing

Compared to the broader internet, Wikipedia is simpler to preprocess due to the rigid formality of the archival process, the MediaWiki markup language, and enforced community standards. For most tasks, this approach is able to extract names, timestamps, and labels with only regular expressions.

Extracting Timestamps: AfD discussion norms require that all contributions are signed using a standard format, which includes the contributor's username or IP address and a timestamp in UTC

format. These signatures are highly formulaic and easy to extract, because they can be automatically generated by MediaWiki's ~~~~ shorthand. When users do not sign contributions, bots add them, along with a citation to the `SIGNATURES` policy. All lines following the outcome are checked for timestamps in Wikipedia standard format.

```
\d\d:\d\d, \w+ \d+, 20\d\d (UTC)
```

In regular expressions, `\w` matches any letter and `\d` can match any numeric character. A `+` suffix captures one or more consecutive characters of that type.

Extracting outcomes: AfD discussions are archived in a specific format with only minor variation, and can be easily extracted for structured representation. I define a discussion as having an *outcome* if its archival page includes a header line with one of three fixed phrases (ignoring whitespace):

```
The result of the debate was [x]
The result was [x]
The result of this discussion was [x]
```

This pipeline saves the captured string `[x]` as the debate outcome. When these lines are timestamped, the user and timestamp of the outcome are also logged as available metadata for analysis.

Extracting nominations, votes, and comments: If a timestamped contribution appears at the top of the discussion, prior to any votes, it is treated as a *nomination*. These statements have become more common over time: while they occur in only 67% of nominations in 2005, they were rapidly adopted and are present in 98% of nominations since 2008; under present policy, omitting a nominating statement is an acceptable reason for "speedy" dismissal and default "Keep" outcome for an AfD nomination.

Following the nominating statement, any timestamped line is captured as either a vote or a comment. I define votes as any timestamped line beginning with a bolded phrase, following Wikipedia convention for contributions:

```
* "'[y]"'
```

Posts beginning with one or more leading asterisks creates a bulleted, threaded discussion. Words or phrases surrounded with three apostrophes creates '''**bolded**''' text. The value of this bolded text `[y]` is captured and stored. If no bolded phrase is present, but the line is still signed and timestamped, that line is treated as a *comment*. Lines beginning with the bolded phrase **"Comment"** are also treated as comments. Lines beginning with **"Note"** are automatically generated, typically for categorizing discussions by topic, and are discarded. Lines with "Relist" bolded are administrative notes to keep the discussion open for longer than the typical seven days, and are also discarded. Lines with no timestamped signature are discarded.

Several alternative solutions to deletion exist; each maintains the content of the page while deleting the page itself. In the five-label case, `Merge` and `Redirect`, the two most common alternate outcomes, are represented separately in line with prior work; in the two-label case they are merged in with Delete. All other values are grouped together as `Other` in the five-label case ("Userfy", "Transwiki", "Move", and "Incubate"). In the two-label case they are merged in with `Keep`. Votes and outcomes of "Close", "Withdraw", and "Cancel" are treated as "Keep" outcomes as the page as well as its content is fully maintained. Copyright violations are treated as a "Delete" outcome, as the content is deleted as a result of the outcome. Any given vote or outcome is represented as a set that can contain zero or more normalized labels. Therefore, the probability of a vote for a particular label is not drawn from a distribution; probabilities of each label in L are disjoint.

Extracting users: For each nomination, outcome, vote, or comment, I log the user whose signature immediately appears before the timestamp, either with a MediaWiki link to their User page or their User Talk page:

```
[[User Talk:[z]
[[User:[z]
```

I extract `[z]` as a username and associate it with the nomination, outcome, vote, or comment where it was captured. When user signatures link to both User and User Talk pages and those usernames differ, the Talk page's username is prioritized.

The public release of this corpus includes designated fold assignments for reproducible results and future comparisons against baselines on the 5% subset used in this work. It also includes two formats for experimenting with the full corpus: a 10-fold cross-validation split, as well as a single train/validation/test split for use with more resource-intensive classifiers, especially neural methods. The library that I developed for producing these variables is written in Python and compatible with standard implementations of BERT and a standard JSON format for representing group discussions. After publication of the papers associated with this work, I also released an update to make the corpus compatible with standard Pandas dataframes.

The public release of this data includes the full corpus, including the 8.1% of filtered nominations with no discussion or missing outcomes; labels for all votes and outcomes, in three levels of granularity. The released corpus can be normalized to a two-label model using only `Keep` and `Delete`, or to a five-vote model that also maintains separate categories for `Merge`, `Redirect`, and `Other`. The corpus also preserves raw text of votes and outcomes, which include a very long tail of free-form inputs. For all analyses in this work, I use the

two-label model, but the released source code includes options for using the five-label variant. Finally, I have released the manually constructed taxonomy of policies, guidelines, essays, and aliases, and a sample of the format used for this data in JSON format.

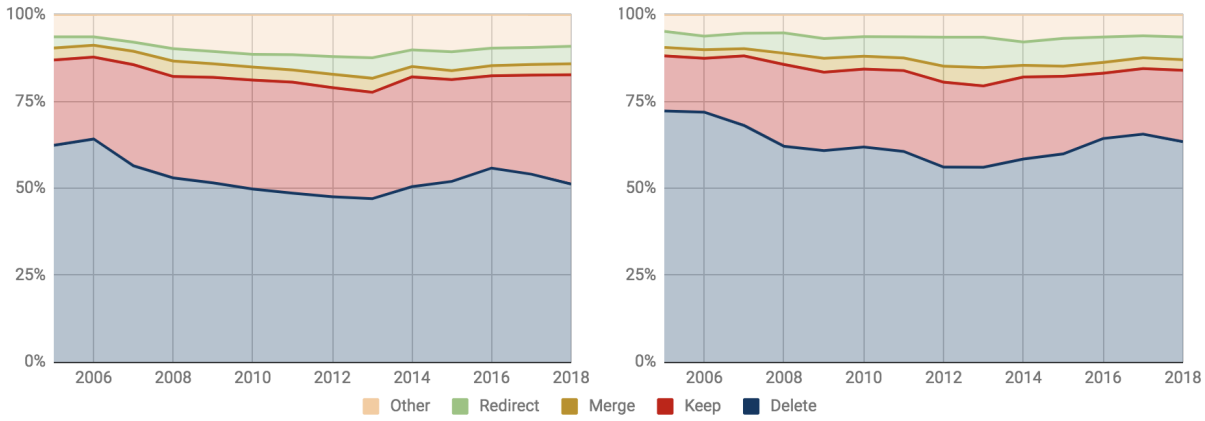


Figure 9: Distributions by year for votes (left) and outcomes (right) over Wikipedia's history.

Corpus Overview

This corpus enables the first comprehensive review of activity statistics in *AfD* since 2010[180]. Figure 9 shows distributions of voting preferences over time, separating votes and outcomes. Vote totals approximately match reported distributions from work early in Wikipedia's history; however, I find a much narrower spread between `Delete` and `Keep` votes compared to early work. While that work showed a 40-point margin in favor of `Delete` (64% to 24%), I only observe a 26.5-point margin. Additionally, I measure distributions of final administrative *outcomes*, and find that outcomes are more deletionist than votes, with `Keep` comprising over 28% of votes but fewer than 21% of final outcomes. Part of this is driven by the increased length and controversy of discussions that lead to `Keep` outcomes - more votes are cast per debate than in uncontroversial `Delete` decisions. Additionally, the presence of long-tail rare labels is more common in outcomes than in votes, such as references multiple outcomes ("`Merge and Delete`," for instance), or outcomes resulting in `No Consensus` (which defaults to a `Keep` outcome, functionally).

This gap is partially explained by the difference in time period observed in our dataset. `Delete` votes were already becoming less common in the later years of that study's window of observation, a pattern that has since been maintained. The decline in site activity was linked to a continued decrease in `Delete` votes, falling from a peak of 64.1% in 2006 to a low of 47.0% in 2013, then seeing a modest

[180] Dario Taraborelli and Giovanni Luca Ciampaglia. "Beyond notability. Collective deliberation on content inclusion in Wikipedia". In: *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*. 2010, pp. 122–125

resurgence but mostly stabilizing over the last decade at levels lower than the early peak.

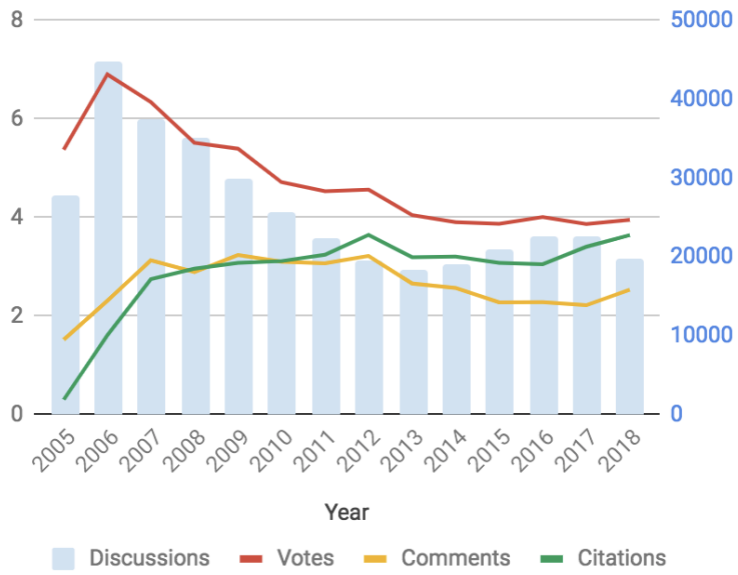


Figure 10: Counts of discussions per year (blue) and of votes, comments, and citations *per discussion* in each year.

Mirroring the overall drop in editor activity over time, voting activity in debates has declined over time. After reaching a peak of 6.9 votes per discussion in 2006, activity declined, and discussions have averaged 4.3 votes in the past ten years. Figure 10 gives volume counts for discussions over time. In contrast to the decline in voting activity, the data shows a slow and steady growth of citation to policy. In early years of the site, votes outnumbered citations to policy by a ratio of more than ten to one. In the most recent year, vote and citation counts are near parity.

Throughout this analysis, I also measure success rates for votes, defining a user's vote as successful if its label matches the outcome decided by an administrator. Across all votes in AfD's history, 67.9% have been successful (matching the final outcome of the debate). This number rises to 75.6% when only considering votes for `Keep` or `Delete` outcomes. Overall, deletionism is a "safer" bet: `Delete` votes are successful 82.0% of the time, while `Keep` votes have a 64.0% success rate.

User Distributions

The full set of contributors to our corpus is made up of over 161,266 editors in a log-normal distribution; by log-likelihood ratio, log-normal more closely fits contribution counts than other heavy-tailed distributions, $p < 0.01$. This distribution is visualized in Figure 11.

Half of all contributions are made by 1,218 users, or just under 0.8% of editors present in our corpus. In contrast, 124,826 observed users (77.4%) contributed fewer than 5 edits; cumulatively, they account for only 5.7% of the observed data. Most frequently, users enter *AfD* to participate in a single debate, and never return. These results replicate the observation from early work by Schneider et al.[181] and Geiger & Ford[182] that *AfD* is dominated by long-time members rather than newcomers; in fact, as this trend has increased in recent years, the distributions observed are *more* extreme than what has been previously reported.

With this corpus in hand and these initial analyses as a context and sanity check on the dataset, we are now ready to move on to our machine learning tasks in the domain.

[181] Jodi Schneider, Alexandre Passant, and Stefan Decker. "Deletion discussions in Wikipedia: Decision factors and outcomes". In: *Proceedings of WikiSym*. ACM. 2012, p. 17

[182] R Stuart Geiger and Heather Ford. "Participation in Wikipedia's article deletion processes". In: *Proceedings of WikiSym*. 2011, pp. 201–202

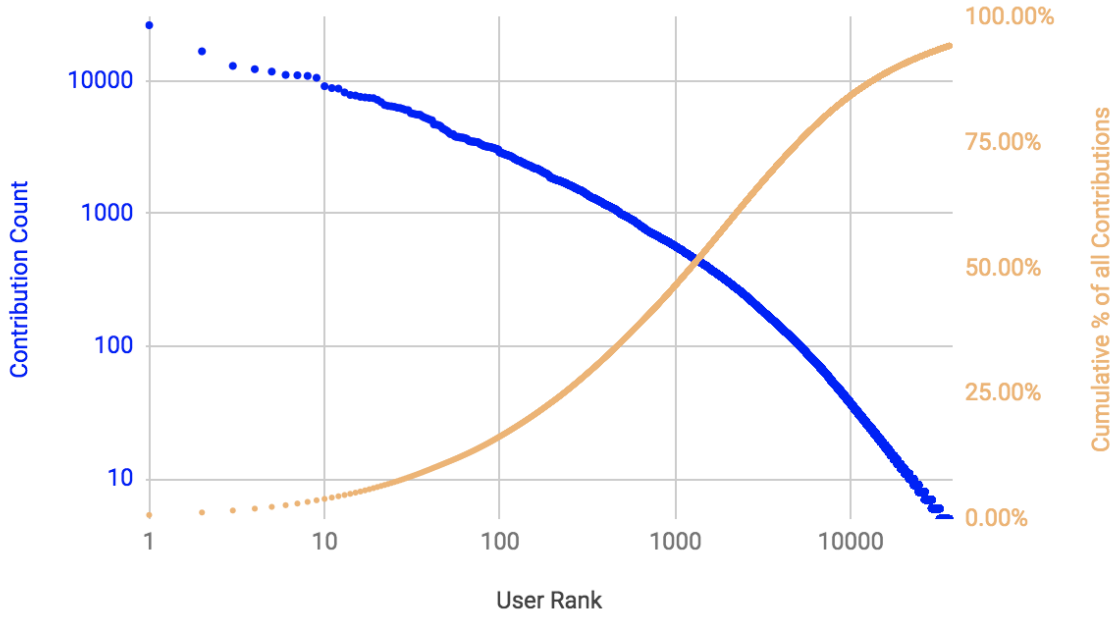


Figure 11: Log-log plot of user rank and contributions. The top 36,440 users, all with at least five contributions, are displayed. Collectively, these 22.6% of all users account for 94.3% of all contributions.

Predictions in Wikipedia

I'll be modeling the discourse of *AfD* with two classification tasks for studying group decision-making processes.

- **Stance classification**, a fine-grained, fully supervised classification task for individual contributions to a discussion.
- **Outcome prediction**, a distantly supervised task requiring far less annotated training data for new domains.

In other fields, the term "preference" is often used where NLP researchers would say "stance." Throughout this work, I use these terms mostly interchangeably. To begin, I'll demonstrate that these tasks are tractable for NLP researchers today, especially with modern language representations like BERT[183]. This contextual representation is highly accurate in both supervised tasks and produces interpretable results for the unsupervised task, suggesting it is ready for immediate application in social science research.

[183] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL*. 2019

Throughout the analyses to follow, I use the following notation:

- A single deletion discussion is labeled d. It has a series of contributions $[c_0, c_1 \dots c_N]$.
- Each contribution c_i has a corresponding username u_i, vote label l_i (null for comments), timestamp t_i, and a rationale text, r_i (which might be empty).
- From a discussion d, a machine learning classifier extracts representative features ϕ and produces a posterior probability distribution $P(l|\phi)$, where the total probability of all labels sums to 1. The features of a discussion at the moment contribution c_i was posted (all comments up to timestamp t_i) are represented as ϕ_i.

In the corpus as released for public use, I provide two possible labeling schemes L, a 2-label case for binary classification, used for new experiments and analysis, and a 5-label case for direct comparison with prior work like Lam et al.[184].

[184] Shyong K Lam, Jawed Karim, and John Riedl. "The effects of group composition on decision quality in a social production community". In: *Proceedings of Group*. ACM. 2010, pp. 55–64

$L_2 = \{\texttt{Delete}, \texttt{Keep}\}$

$L_5 = \{\texttt{Delete}, \texttt{Keep}, \texttt{Merge}, \texttt{Redirect}, \texttt{Other}\}$

Machine learning is performed as described in Chapter 2, with text representations varying between a bag-of-words model, a *GloVe* dense embedding, and a *BERT* contextual embedding. The *BERT*

model was already trained on Wikipedia texts (and other sources), so I perform no fine-tuning. This may mean text from thecorpus is included in $BERT_{BASE}$ training data, causing a minuscule exposure to test data in my experimental setup; I do not investigate this question here, but note it as a complicating factor. Experiments represent average results of 10-fold cross-validation. All instances from a particular discussion appear in only one fold; there is never crossover from the same debate between train and test data. I report results on a randomized subset of 5% of the corpus, approximately 20,000 discussions. In preliminary evaluation, a 20x growth in training data increased computational resources beyond what is practical for social scientists, for model accuracy improvements of less than 1%; I provide training splits (for potential future approaches that benefit from larger corpora) in the released corpus.

Turn-Level Stance Classification

In most other collaborative team decision-making contexts, opinions are expressed but explicit stances are latent. Because of the unique format of Wikipedia discussions, those stances are easily extracted from **'''bolded'''** votes. I use this as a test case for building supervised classifiers which elicit participant stance based on their statements alone. All bolded text is masked from rationales and models must predict what vote is associated with a given rationale.

Fundamentally this is a test of how closely the Wikipedia domain hews to other decision-making contexts. If rationales are *not* sufficient to predict stances accurately, it means one of two things. Either rationales do not carry information about user preferences, and so are not comparable to group decision-making in contexts where those preferences are not explicitly labeled with votes; or the rationales do carry this information, but they are not tractable with current NLP methods. To evaluate this, I define a task to label each vote in each *AfD* discussion:

- **Possible Labels:** $L = \{\texttt{Delete}, \texttt{Keep}\}$
- **Input:** Rationale text r_i from a single vote.
- **Features:** A representation vector $\phi(c_i)$.
- **Output:** A predicted stance $l \in L$.

I exclude non-voting comments from this analysis, as no gold labels are available for supervised training. Expansion to distant supervision, where user stances from votes are used as gold labels for that user's comments, is a possibility for future work. User stances

Representation	Accuracy	
	%	κ
Majority Class	63.8	0.00
GloVe	76.0	0.45
Bag-of-Words	81.8	0.59
BERT	82.0	0.60

Table 5: Accuracy of stance classification models for individual contributions, based on rationale text alone.

are explicitly given by users in the original corpus and there is no ambiguity; the upper bound for this task is 100% accuracy and $\kappa = 1.0$. Individual votes or comments have short rationales, however, typically only a sentence or a few words.

Despite this, n-gram models provide a robust baseline, and while the BERT model outperforms a unigram baseline, the difference is small. Comparing embeddings, the newer contextualized BERT model outperforms GloVe by more than 6% absolute and 10% relative. Overall, this result shows that the stance classification task is tractable, with good accuracy. The positive result mirrors similar tasks have been effective in labeling turns in prose text (see work[185] by Wilson et al. and other work with their MPQA corpus); in open-ended group dialogues[186] (including my own prior work[187]); and in stance classification for more open-ended contexts like social media[188]. The result gives a useful proof-of-concept that text rationales carry recognizable stance information and can be reliably classified by a machine learning model.

[185] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis". In: *Proceedings of EMNLP*. 2005

[186] Andreas Stolcke et al. "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: *Computational linguistics* 26.3 (2000), pp. 339–373

[187] Jin Mu et al. "The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions". In: *International journal of computer-supported collaborative learning* 7.2 (2012), pp. 285–305

[188] Parinaz Sobhani, Diana Inkpen, and Stan Matwin. "From argumentation mining to stance classification". In: *Proceedings of the Workshop on Argumentation Mining at NAACL*. 2015, pp. 67–77

Outcome Forecasting

The stance classification task above has limitations for practical use in other group decision-making research. Foremost, it requires training data with labeled votes; this is difficult to get in many cases. Moreover, the stances of individual votes in a discussion are too granular for process variables that aim to represent discussion dynamics overall.

A more relevant goal for social scientists is analysis of group discussions where the preferences of individuals are *unlabeled*, even in training data. Next, I aim to predict the consensus preference of a group, after discussion. This task measures whether language representations can model the many turns in a discussion and mimic the behavior of administrators. To do this, I give as input the rationale texts of nominations, votes, and comments throughout a discussion, and treat the label from administrative closure of a debate as the *only* supervised label of group consensus.

- **Possible Labels**: $L = \{\texttt{Delete}, \texttt{Keep}\}$

- **Input**: Discussion d, with nomination c_0, followed by votes and comments $c_1 \dots c_N$.
  Each contribution c_i consists of:
  - User ID u_i.
  - Timestamp t_i.
  - Rationale text r_i.
  - Stance label $l_i \in L$, or for comments, $l =$. In experiments other than gold-label comparison, l_i is masked.
- **Features:** A representation vector $\phi(d)$.
- **Output**: An outcome label $l \in L$.

For text embedding, I again extract features ϕ_{GloVe} and ϕ_{BERT}, but in this case there is a need to combine vectors from multiple contributions $[c_0, c_1, \dots c_N]$ into a single vector for discussion d. To do so, I encode each contribution's rationale r_i separately (again removing all occurrences of **'''bolded'''** text to mask votes). I then average each contribution's vector, normalized for length:

$$\phi(d) = \frac{\sum_{i=0}^{N} \frac{\phi(c_i)}{\ln(len(r_i))}}{N}$$

Unlike in the first task, outcome prediction is distantly supervised and the task is sometimes undecidable; as discussed previously, administrators occasionally close conversations with results of `No consensus`. To evaluate an upper bound on model accuracy with masked preferences, I include a gold feature vector $\phi^*(d)$ where gold-standard user preference labels *are* made available for modeling. Specifically, for each possible $l \in L$, this vector includes the raw count and percent of votes that label received. While Wikipedia is not a direct democracy, administrators rarely deviate from consensus; this represents a good approximation of an upper bound on meaning representation from rationales alone.

I first train a machine learning model to forecast the outcome of debates from observable characteristics. This model is an estimate of the expressed preferences of a group; the underlying challenge for machine learning is to model the text from many turns in a discussion, and mimic the consensus-forming judgment calls of administrators. For each possible vote label, I extract features including the raw count and percent of votes for that label. While Wikipedia is not a direct democracy, administrators rarely deviate from vote counts, so these vote counts are highly informative. In addition to voting features, I again encode the rationale of each contribution from each

discussion separately using the method described above, then average across the contribution vectors, normalized for length:

$$\phi_i = \frac{\sum_{j=0}^{i} \frac{\phi(c_j)}{\ln(len(r_j))}}{i}$$

The representation of full discussions is simply the special case where all contributions are included, that is, $i = N$.

Importantly, in this case the actual vote is stripped from the text of the rationale. Model performance varies by a large margin depending on whether explicit votes are accessible to the model or not; specifically, the model given access to stances of group members is able to predict outcomes with a Cohen's $\kappa = 0.83$ for full discussions. The BERT model also reaches good levels of agreement, outperforming other representations by at least 1.6% accuracy, in absolute terms. Short discussions are more predictable, with the best-performing model reaching accuracy of 97.3% for short discussions of 5 or fewer total contributions that resulted in a `Delete` outcome, compared to 85.3% accuracy for long discussions of more than 10 contributions that resulted in a `Keep` outcome.

I then test the ability to make incremental predictions, observing only early contributions to a debate. Models are trained identically in this set of experiments; however, in the test set, I create a new instance for classification after *each* contribution to each discussion. Note that reported accuracy in this setup overweights more contentious debates - with more contributions, there are more instances from that discussion to classify in each test set. This slight bias results in over-representation of debates that ended in a `Keep` outcome, as those debates tend to have more contributions, and therefore increases the difficulty of the problem (`Keep` is a minority label and more challenging to predict). In this evaluation, all models see significant performance degradation, with lower accuracy from forecasting early in the debate. GloVe and bag-of-words models are more competitive, but BERT maintains the highest accuracy, with an overall accuracy of 79.7% across all instances and $\kappa = 0.55$.

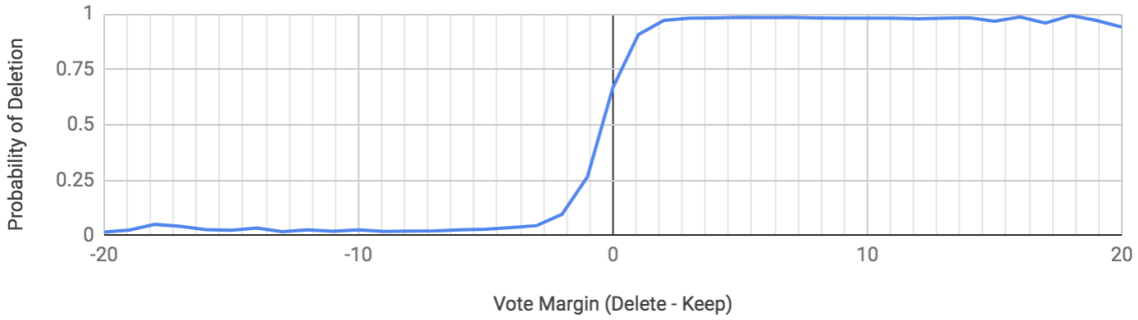


Figure 12: Probability of a `Delete` outcome as voting margin varies. Administrators almost never overrule `Delete` majorities with a margin of at least 2 votes, or `Keep` majorities with a margin of at least 4 votes.

Let's pause quickly to ask: why are models with direct access to votes so accurate? As shown in Figure 12, this result is unsurprising given actual behaviors by administrators. Only 7.6% of votes end in ties; administrators choose `Delete` in 66.9% of these cases). Outside of ties, administrators follow the majority vote in 94.8% of discussions. But while the forecast model that takes language into account does not differ in the accuracy of the model using only gold labels, the text input is highly granular for quantitative analysis tasks that study individual contributions. From now on, I use the model that is given access to all observable information at training time, including both the gold labels and text of individual contributions.

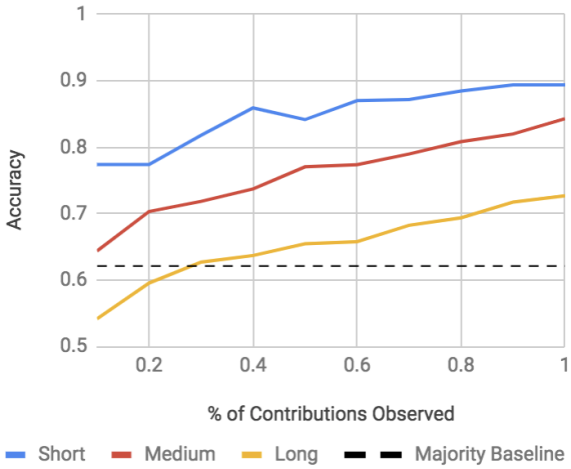


Figure 13: Real-Time BERT model accuracy mid-discussion, split by final debate length: short (5 or fewer), medium (6-10), and long (over 10).

Table 6 shows the full comparison of models. As expected, the model given access to stances of group members is highly accurate. That model is able to predict outcomes with a Cohen's $\kappa = 0.84$ for full discussions. The BERT model also reaches good levels of agreement, outperforming other language representations by at least 1.6% absolute. In the real-time evaluation, GloVe and bag-of-words models are more competitive, but BERT maintains the highest accuracy. All models (including the gold-standard) see significant performance degradation, suggesting that discussions are *not* foregone conclusions after early contributions. To demonstrate this more clearly, see Figure 13, where in conversations of any length, outcome prediction early in the debate is less reliable, then improves in accuracy steadily over time as more contributions are made visible to the classifier.

Error analysis shows that on top of support for social sciences, the remaining errors in classification are an opportunity for improved NLP methods. For instance, in stance classification, there are some cases where individual contributions simply lack the content that is necessary to classify them accurately (e.g. *"Per all the above."*). These

Representation	Full Debate		Incremental	
	%	κ	%	κ
Majority Class Baseline	74.0	0.00	62.1	0.00
GloVe	81.7	0.49	69.1	0.31
Bag-of-Words	84.2	0.58	72.4	0.39
BERT	85.8	0.62	73.4	0.41
BERT + Vote Labels	93.5	0.83	79.7	0.55

Table 6: Accuracy of forecasting for full discussions and incremental predictions.

cases would benefit from a more detailed awareness of threads of conversation[189]. Even more often, classification errors occur when users *themselves* express uncertainty:

> (voting for `Delete`) *"[...] as I said, I am not really qualified to assess these sources in a deeper way, other than to indicate their existence, and "apparent" reliability under our usual sourcing guidelines."*

[189] Justine Zhang et al. "Characterizing online public discussions through patterns of participant interactions". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–27

Instances like these require not just classification for stance but also for uncertainty[190]. Multi-task learning is a particularly fruitful domain for neural methods and the public release of the full corpus should be a resource for development of that field.

[190] Kate Forbes-Riley and Diane Litman. "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor". In: *Speech Communication* 53.9-10 (2011), pp. 1115–1136

In outcome prediction, text-only models underperform the gold-labels model when predicting an outcome of `Keep`, particularly for short debates. As seen in Table 7, when predicting `Delete` in short discussions, the BERT model is almost always accurate; as conversations grow, `Delete` predictions become less reliable, at just over 75% for debates longer than 10 contributions. By contrast, when BERT predicts `Keep` it becomes *more* accurate as conversations grow. In short discussions where the final outcome was `Keep`, performance is at its worst, with a gap in accuracy over 22% compared to the gold model. This suggests that there is significant opportunity to better identify *persuasive* early `Keep` votes, which are elusive in existing representations. Further technological advances may also focus on recognizing short discussions that *ought* to be enhanced with additional evidence, either through intelligent routing to potential participants or direct intervention with relevant content.

Final	ϕ	Short	Medium	Long
Delete	BERT	92.9	85.6	74.7
	Gold	97.3	92.9	85.4
		(-4.4)	*(-7.3)*	*(-10.7)*
Keep	BERT	71.9	80.6	75.0
	Gold	91.8	92.2	85.3
		(-19.9)	*(-11.6)*	*(-10.3)*

Table 7: Accuracy of outcome prediction, split by final outcome and total debate length (as in Figure 13).

Explaining Wikipedia Decisions

In the chapter that follows I now put these predictive models to work. Let's see what we can learn about this domain by leaning on the trained classifiers from the previous chapter.

As I discussed in the first portion of this thesis, lack of causality is frequently cast as a limitation in studies of online communities, certainly including *AfD*. Similar limitations applied to the early analysis Taraborelli & Ciampaglia[191] that showed a "herding" effect, where later votes in a discussion tended to follow early votes; this too may have been because votes are an accurate proxy of article quality rather than a rhetorical impact of early voters on those who participate later. Notably, the studies that have focused on *rhetoric* in debates have avoided tying these analyses to success rates. For instance, Xiao & Sitaula measured sentiment of votes and found more positive affect in `Keep` votes, but did not correlate sentiment to outcomes[192]; similarly, Schneider et al. investigated the rhetorical argument strategies of editors in *AfD* but did not measure how those strategies affected success rates or influenced future decisions[193]. But as I have argued, causality is not necessary to effectively describe the constraints of the Wikipedia *AfD* domain. In the next section I demonstrate the explanatory value of these models even when causality is not part of an analysis.

[191] Dario Taraborelli and Giovanni Luca Ciampaglia. "Beyond notability. Collective deliberation on content inclusion in Wikipedia". In: *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*. 2010, pp. 122–125

[192] Lu Xiao and Niraj Sitaula. "Sentiments in Wikipedia Articles for Deletion Discussions". In: *International Conference on Information*. Springer. 2018, pp. 81–86

[193] Jodi Schneider et al. "Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups". In: *Proceedings of CSCW*. ACM. 2013, pp. 1069–1080

Forecast Shifts

I measure shifts in probability output from the forecast model at each of these incremental predictions. I measure the change in the probability distribution of outcomes immediately after each contribution is posted. For nominations ($i = 0$), for each possible $P(l)$, for $l \in L$, the prior probability distribution of all outcomes $l \in L$ as measured from training data is subtracted instead.

$$\Delta(l, c_i) = P(l|\phi_i) - P(l|\phi_{i-1})$$

Increase in forecast probability of one label shifts that label upward, and another simultaneously downward, doubling the cumulative impact of changes; therefore, the change in probability for labels is summed then multiplied by a normalizing factor of $1/2$ to produce a measure of *forecast shift*, ranging from [0,1] per contribution.

$$\text{Forecast Shift} = FS(c_i) = \frac{1}{2}\sum_{l \in L} |\Delta(l, c_i)|$$

For corpus study, I measure forecast shift of contributions through 10-fold cross-validation. For each fold, I train a model on the training set, then make incremental probability forecasts using that model for each discussion in the test set. After applying this method across each fold, I am able to measure forecast shifts for the entire corpus, with no discussion received forecasts from a model where that discussion was itself part of the training set.

This approach to explanation has *significant* limitations for making causal claims about the data in the corpus. This measure of forecast shift is a descriptive measure of how a predictive model alters its prediction based on new evidence. However, I have only inspected what is *predictive* given limited information, but not what is rhetorically *influential* to the debates themselves. Thus while the features identified as shifting forecasts are informative, this explanatory work does not make causal claims. As a crucial example of where this limits analysis, I do not include article texts themselves in this work. The predictive model has found that debate-initial `Keep` votes are predictive of `Keep` outcomes. This has at least two possible explanations. The first is that early `Keep` outcomes are persuasive or influential in the debate itself and lead to articles being preserved. The second is that articles worth preserving *attract* early `Keep` votes, and the rhetorical strategy of the voter is unimportant compared to the voter as proxy measure of article quality.

Explanations from Forecast Shifts

Within the analyses that follow, I will begin each section with a reference to the key prior work that informs a particular question. Along with measuring any particular user behavior, like arriving early or posting frequently, I will also point out specific *policies* that are often cited in exemplar cases of that behavior. This is a useful analytic lens; policies have consistently been a focus area of *AfD* research, from the close study of the `Ignore All Rules` policy[194] to the study of notability subpolicies[195]. But in this work, I find the relationship between policies, success, and forecast shift is nuanced, and rather than treating policy citation as a monolithic phenomenon, I choose to name specific policies as they become relevant to other aspects of editor behaviors.

For instance, I find that there is no overall correlation between the success rate of votes in which a policy has appeared, and the mean forecast shift from those votes; the slope is in fact slightly negative

[194] Elisabeth Joyce, Jacqueline C Pike, and Brian S Butler. "Rules and roles vs. consensus: Self-governed deliberative mass collaboration bureaucracies". In: *American Behavioral Scientist* 57.5 (2013), pp. 576–594

[195] Lu Xiao and Nicole Askin. "What influences online deliberation? A Wikipedia study". In: *Journal of the Association for Information Science and Technology* 65.5 (2014), pp. 898–910

($r = -0.04$). Many successful policies do not tend to appear in contributions that changed the model's forecasted outcome, and many contributions that changed the model's predicted outcome dramatically do not end up on the winning side of debates. While this disconnect is important for investigation, it also serves as a limit on the strength of the statements I can make about these findings.

Early Voters

I first evaluate whether early contributions are correlated with discussion outcomes, following on observation of a "herding" effect in early work[196]. In that work, the authors found that later votes in discussions were more likely to mirror early votes. My results replicate this finding: early votes are highly predictive of outcomes. Debate-initial `Delete` votes are successful 84.5% of the time compared to a 63.9% baseline. The effect is even greater for early `Keep` votes, resulting in a `Keep` outcome 62.2% of the time, compared to a 20.7% baseline.

[196] Dario Taraborelli and Giovanni Luca Ciampaglia. "Beyond notability. Collective deliberation on content inclusion in Wikipedia". In: *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*. 2010, pp. 122–125

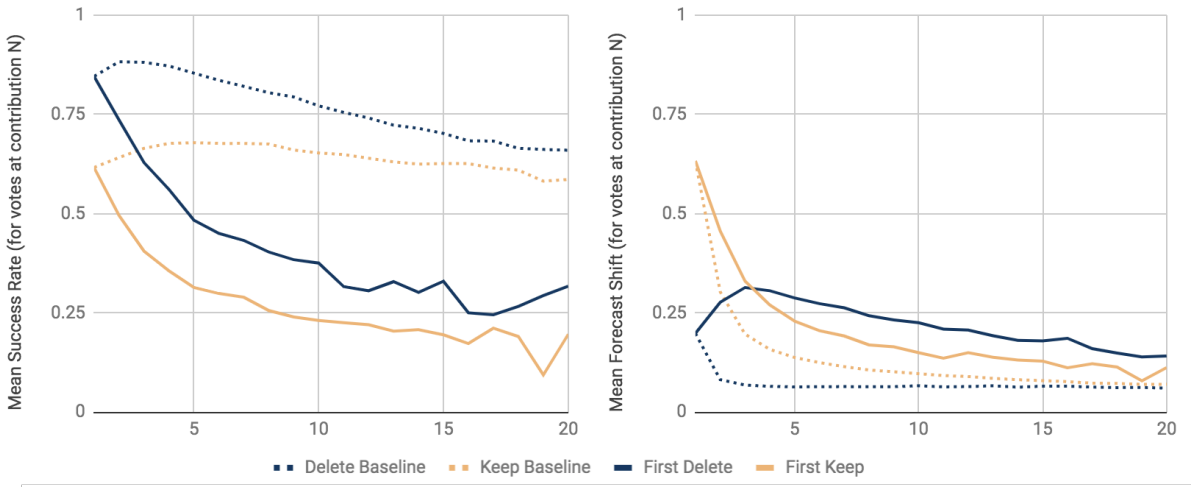


Figure 14: Success rates (left) and forecast shifts (right) for votes that were the Nth contribution to a discussion, for different values of N. I measure these values first for *any* vote with that label at that ordinal location in the debate, then for discussions where the *first* vote for a particular label appeared at rank N.

Trends over the course of a discussion, for both outcome success and forecast shift, are visualized in Figure 14. I separate the values for *all* votes that appear as the Nth contribution to a discussion from contributions at that point that were the first vote for a particular outcome. Success rates rapidly decline for both `Delete` and `Keep` when they arrive late in a discussion. When measuring forecast shift of votes, similar declines are observed for votes overall, regardless of whether they are for `Delete` or `Keep`. Differentiation appears in forecast shift associated with the first `Keep` and `Delete` vote in a discussion. Early `Keep` votes are highly informative for the forecast model, and produce the greatest shift in forecast probabilities. But

forecast shift from `Delete` votes initially *increases* as it appears later in the discussion, peaking at the third contribution and only declining slowly when the first vote appears even later than that.

These results are intuitive. The default outcome when discussions do not reach consensus is `Keep`; however, the momentum in *AfD* is toward deletion. For inclusionist voters, a key factor in highly predictive votes is simply being *early* to arrive in a debate, either shifting the tenor of the discussion that follows or signaling clear article quality or meeting criteria for inclusion. When voting `Delete`, on the other hand, forecasts do not shift when they arrive early; a `Delete` outcome was already likely. Instead, `Delete` voters shift forecasts when they arrive in the middle of conversations and *contradict* earlier votes. The `Delete` voter shifts forecasts more significantly when acting as a "devil's advocate" and *reducing* certainty of a particular outcome; this is not possible in debates where deletion is obvious, and those `Delete` votes result in low values of forecast shift.

Research has shown that priming effects are able to shape risk profiles, preferences, and topics under scrutiny in decision-making tasks[197,198]. In that context it makes sense that early votes can set the stage for later discussion, and that this impact is larger when the first vote is contrary to the most common outcome (`Keep` initial votes are more influential than `Delete`). The activity of late-comers to discussions, adding votes even though they have no impact on discussion, is also a result that is validated and justified by prior work.

[197] Hans-Peter Erb, Antoine Bioy, and Denis J Hilton. "Choice preferences without inferences: Subconscious priming of risk attitudes". In: *Journal of Behavioral Decision Making* 15.3 (2002), pp. 251–262

[198] Eric J Johnson et al. "Beyond nudges: Tools of a choice architecture". In: *Marketing Letters* 23.2 (2012), pp. 487–504

An example of this pattern is shown in Figure 15, where one user defend a page for a sparsely populated island in the Indian state of Kerala. In the figure, I omit the (lengthy) discussion; to summarize, the first voter produces a large forecast shift, beginning the debate with an initial `Keep` vote only two hours after nomination. When later users argue for `Delete`, the model shifts back to predicting a `Delete` outcome, but with low certainty. The next non-voting comment from the initial `Keep` voter gives detailed responses and further citation to policies, which tilts the forecast toward an eventual `Keep` outcome.

Notability Policies

A differentiating feature of the voter in the previous example is the citation of relevant and targeted policy, `Notability (geographic features)`. This type of behavior was previously flagged in preliminary results from prior work from Xiao et al.[199], which noted that locations, biographies, and corporate pages were each deleted at different rates compared to pages in general. My research extends that finding: Notability policies are among the most informative votes

[199] Lu Xiao and Nicole Askin. "What influences online deliberation? A Wikipedia study". In: *Journal of the Association for Information Science and Technology* 65.5 (2014), pp. 898–910

Edayilakkad [edit]

Quality issues, possibly beyond any fixing. See discussion at talk:, and at Wikipedia:Village_pump_(proposals)#I.27ve_had_enough._.22Approved_articles.22_clearly_no_better_than_ones_that_skip_it Andy Dingley (talk) 00:11, 7 July 2017 (UTC)

- **Keep**. Per WP:GEOLAND an inhabited island is presumed Notable, and in my opinion that is an almost automatic qualification for inclusion if meaningful information can be verified. While I haven't yet found WP:GNG's usual expectation for significant coverage in any particular source, I have been finding a fair number of sources with various brief mentions. Note that source searching is difficult because there are several variations on the spelling, and because useful search results tend to be heavily buried under garbage search results. Alsee (talk) 02:35, 7 July 2017 (UTC)

  P.S. An inhabited island in the U.S. would almost certainly be kept, and if this is deleted I'm sure it would just get re-created in a few years as India comes more online with sources. Alsee (talk) 03:09, 7 July 2017 (UTC)

  P.P.S. Here's the article at Malayalam language Wikipedia: ml:ഇടയിലക്കാട്. There are a mix of blog-sources as well as usable sources in that version. Alsee (talk) 04:11, 7 July 2017 (UTC)

  > It's not about whether an island is implicitly notable, it's about whether this article passes our standards to adequately demonstrate that. WP:RS and WP:V are strong policy. WP:OTHERSTUFFEXISTS is not. Andy Dingley (talk) 09:11, 7 July 2017 (UTC)
  >
  > > Andy Dingley, I do not disagree with your concerns about quality. However the excessive 11 keeps here indicate that you've missed a significant detail. Your first sentence got it backwards, it is about whether the island is implicitly Notable. Notability isn't a property of the article, it's a property of the topic. An article that contains zero evidence of notability is a Keep, if sources exist and the topic itself satisfies Notability. In the most extreme case you keep the article and delete all the junk down to a single sentence stub. An irredeemably promotional article on a company might get hit with an unsympathetic TNT, but we're going to salvage anything we can for a desirable article on an inhabited place. Documenting significant geography is about as close to objectively-desirable as it gets. Alsee (talk) 23:53, 13 July 2017 (UTC)

Figure 15: Large forecast shifts arise from initial votes for `Keep` followed by response votes for `Delete`. Here, a user successfully cites the `Notability (geographic features)` policy to keep an article.

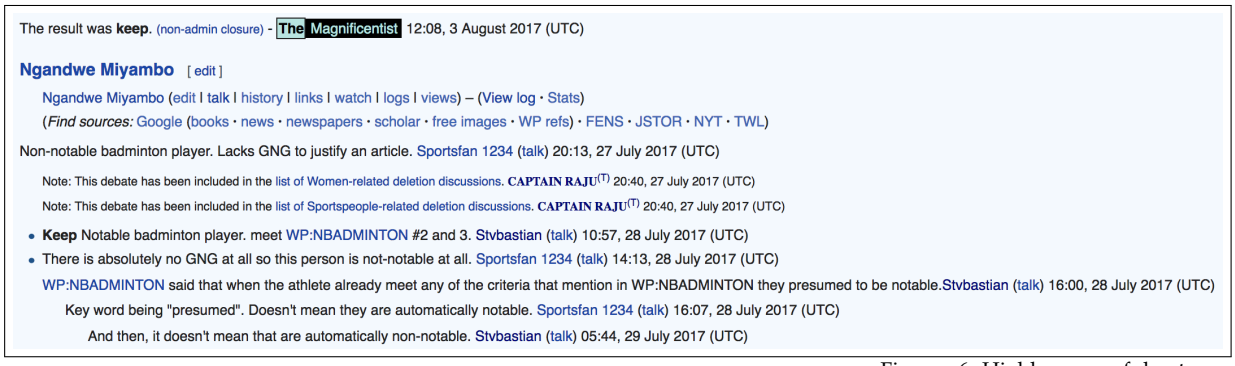
	Success		Forecast Shift	
Notability Subcategory	`Keep`	`Delete`	`Keep`	`Delete`
Biographies of Living People	54.5	89.5	0.31	0.11
Astronomical Objects	78.6	60.3	0.29	0.12
Martial Arts	56.6	92.5	0.27	0.09
Software	41.7	92.6	0.251	0.08
Media	75.9	87.1	0.246	0.06
Films	81.3	84.3	0.237	0.07
Academic Journals	75.2	77.6	0.23	0.04
Professional Football Leagues	82.5	93.4	0.22	0.07
Music	72.1	85.7	0.21	0.08
Geographic Landmarks	85.7	74.5	0.21	0.12

Table 8: Success and forecast shift for Notability citations, split by vote label (`Keep` or `Delete`).

in the forecast model, appear early in debates (particularly often in `Keep` votes), and are more successful in general than other policies and more than votes in general. This is shown in the ranked list of Notability policies associated with the greatest average forecast shift, in Table 8.

Successful policy citations that also have high forecast shift are narrowly scoped. The most successful inclusionist Notability policies are on topics like astronomical objects, geographic landmarks, and local high schools. Enthusiasts wrote these policies to clearly define notability for an area where the average editor may not know inclusion criteria, and cite these policies effectively, first to shift the focus of discussions and then to win those debates. Note that some communities reverse this trend and maintain highly selective standards

to prevent an influx of articles; this phenomenon is most prevalent in sports, with highly successful citations in favor of `Delete` for topics like regional football (soccer) leagues and martial arts. For a prototypical example of highly successful citations, see the actions in Figure 16. This user is one of the top five most consistently successful users in the corpus, by average success and forecast shift. Their contributions are early, short, clear, and uncontroversial. By referencing criteria in a pre-existing, relevant policy (in this case, `Notability (Badminton)`), debate is closed quickly. In fact, the overwhelming majority of this user's votes are for badminton players, all of whom meet the officially written policy's standards, and this user maintains a success rate in excess of 90%.


The result was **keep**. (non-admin closure) - The Magnificentist 12:08, 3 August 2017 (UTC)

Ngandwe Miyambo [edit]

Ngandwe Miyambo (edit | talk | history | links | watch | logs | views) – (View log · Stats)
(*Find sources:* Google (books · news · newspapers · scholar · free images · WP refs) · FENS · JSTOR · NYT · TWL)

Non-notable badminton player. Lacks GNG to justify an article. Sportsfan 1234 (talk) 20:13, 27 July 2017 (UTC)

Note: This debate has been included in the list of Women-related deletion discussions. CAPTAIN RAJU(T) 20:40, 27 July 2017 (UTC)

Note: This debate has been included in the list of Sportspeople-related deletion discussions. CAPTAIN RAJU(T) 20:40, 27 July 2017 (UTC)

- **Keep** Notable badminton player. meet WP:NBADMINTON #2 and 3. Stvbastian (talk) 10:57, 28 July 2017 (UTC)
- There is absolutely no GNG at all so this person is not-notable at all. Sportsfan 1234 (talk) 14:13, 28 July 2017 (UTC)

WP:NBADMINTON said that when the athlete already meet any of the criteria that mention in WP:NBADMINTON they presumed to be notable.Stvbastian (talk) 16:00, 28 July 2017 (UTC)

Key word being "presumed". Doesn't mean they are automatically notable. Sportsfan 1234 (talk) 16:07, 28 July 2017 (UTC)

And then, it doesn't mean that are automatically non-notable. Stvbastian (talk) 05:44, 29 July 2017 (UTC)

Figure 16: Highly successful votes that also shift the forecast model often come from the narrow use of established policies for notability in specific subtopics.

But as `Notability` policies become broader, their trends in both success rates and forecast shifts revert to the broader mean of all votes that cite policies. The very broadly scoped policy on proposed deletion of biographies of living people (`WP:BLPPROD`) is noteworthy: among all policies, it has the greatest difference in success and forecast shift metrics depending on whether it appears in `Delete` or `Keep` votes. When used in inclusionist arguments, the policy is usually cited early in the discussion and causes significant uncertainty in the model, shifting probable outcomes from being weighted toward delete to a tossup. However, those `Keep` citations of the biography policy are among the least successful votes in the corpus. By contrast, when cited as part of `Delete` arguments, this broad policy does much less to shift forecast probabilities, but is successful well above the baseline success rate for deletionist votes. Another way of seeing the role of Notability policies in debate is to look at the `Delete` policy citations with high average forecast shifts. While `Keep` votes have disproportionately high forecast shift values, the top two `Delete` citations in forecast shift are `Trivial Mentions` and `Existence` $\neq$ `Notability`. Both of these policies are used as responses to notability arguments from `Keep` voters.

Taken in aggregate, these results match the findings from Xiao in

her series of papers[200], showing that Notability policies shape the discourse of this domain, while giving substantial additional detail. Notability citations are the most interesting and valuable source of research on *AfD* in future work.

[200] Lu Xiao and Nicole Askin. "What influences online deliberation? A Wikipedia study". In: *Journal of the Association for Information Science and Technology* 65.5 (2014), pp. 898–910

Active Voters

I next evaluate whether votes from more *frequent* posters, who take the time to reply to other users and participate actively in discussion, are more predictive of future outcomes. This effect has been previously suggested in small-scale, mixed-methods analyses of dozens or hundreds of discussions[201,202]. I find mixed support for these findings in a larger-scale study. In the data, 45.0% of votes and comments in discussions come from editors that made more than one contribution in that discussion. Of these, single-contribution voters are substantially more likely to cast a successful vote, winning 84.8% of `Delete` votes and 66.7% of `Keep` votes. Users who post more than twice to a discussion are successful in fewer than half of their votes, and success rates continue to decline as users post more and more. This seems to contradict the topline finding from past work. Again, as with all findings, this result is not causal: there is no way with the data gathered here to discern whether editors with weak arguments tend to add more comments to discussions, or whether their heavy participation in debates *causes* them to lose debates. But in either case, there is no evidence to suggest that active users are more likely to win debates.

[201] Elisabeth Joyce, Jacqueline C Pike, and Brian S Butler. "Rules and roles vs. consensus: Self-governed deliberative mass collaboration bureaucracies". In: *American Behavioral Scientist* 57.5 (2013), pp. 576–594

[202] Lu Xiao and Nicole Askin. "What influences online deliberation? A Wikipedia study". In: *Journal of the Association for Information Science and Technology* 65.5 (2014), pp. 898–910

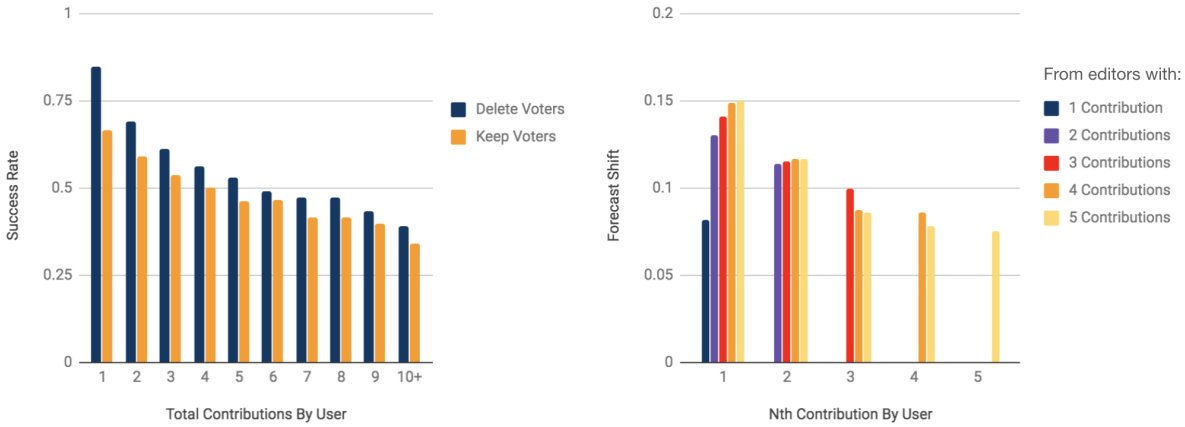
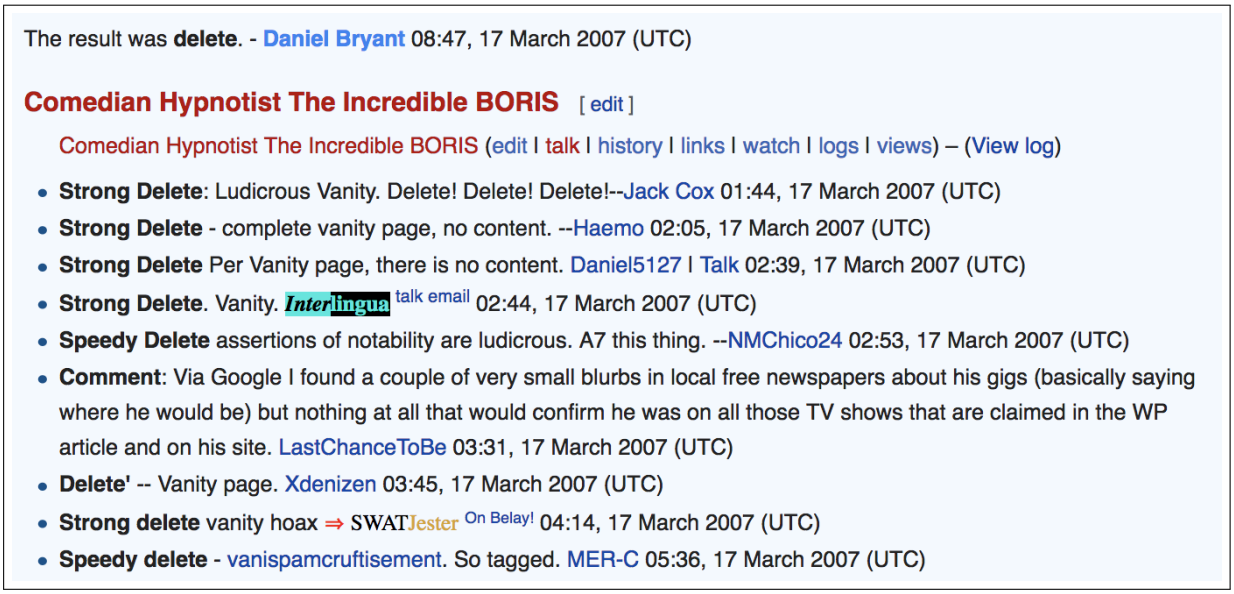


Figure 17: One-time voters are more successful than more active voters; however, the first contribution from more active voters have greater forecast shift than the votes from one-time contributors.

I find that the forecast shift measure matches the smaller-scale observations from those earlier studies, more closely than actual success rates. As shown in Figure 17, while success rates go down as users are more active in debates, the average forecast shift attributable to the first vote from those users is much higher. Forecast shifts are

greater for the first post by editors who will eventually follow up with more activity, though their additional contributions do not maintain that level, suggesting diminishing returns. The first post by these highly active users (the lighter-shaded bars) shifts forecasts by almost twice as much as the first post by one-time contributors (the dark leftmost line).

As an example of this dynamic, see the debate activity in Figure 18, arguing about a Canadian magician. In this debate, several users are successful in their vote, but do not meaningfully contribute to the decision-making process; in the forecast model, only one vote shifts the predicted outcome by more than 0.05, the very first by `Jack Cox`. By the time votes appear from later users, the discussion is a foregone conclusion for `Delete`. The late citation of `Vanispamcruftisement`, a lighthearted anti-spam policy, has no bearing on the clear consensus of the group. While single-vote users are highly successful, they are not changing the outcome of debates; instead, those late arrivals are getting credit for participation in a debate that has essentially concluded. Other citations to policies about spam and hoaxes follow a similar pattern, among the top success-rate policies in the dataset but appearing in votes with among the least new information for the forecasting model.



The result was **delete**. - Daniel Bryant 08:47, 17 March 2007 (UTC)

Comedian Hypnotist The Incredible BORIS [edit]

Comedian Hypnotist The Incredible BORIS (edit | talk | history | links | watch | logs | views) – (View log)

- **Strong Delete**: Ludicrous Vanity. Delete! Delete! Delete!--Jack Cox 01:44, 17 March 2007 (UTC)
- **Strong Delete** - complete vanity page, no content. --Haemo 02:05, 17 March 2007 (UTC)
- **Strong Delete** Per Vanity page, there is no content. Daniel5127 | Talk 02:39, 17 March 2007 (UTC)
- **Strong Delete**. Vanity. Interlingua talk email 02:44, 17 March 2007 (UTC)
- **Speedy Delete** assertions of notability are ludicrous. A7 this thing. --NMChico24 02:53, 17 March 2007 (UTC)
- **Comment**: Via Google I found a couple of very small blurbs in local free newspapers about his gigs (basically saying where he would be) but nothing at all that would confirm he was on all those TV shows that are claimed in the WP article and on his site. LastChanceToBe 03:31, 17 March 2007 (UTC)
- **Delete'** -- Vanity page. Xdenizen 03:45, 17 March 2007 (UTC)
- **Strong delete** vanity hoax ⇒ SWATJester On Belay! 04:14, 17 March 2007 (UTC)
- **Speedy delete** - vanispamcruftisement. So tagged. MER-C 05:36, 17 March 2007 (UTC)

Figure 18: Example of highly successful editor behavior with minimal forecast shift. For each of the later votes, the probability of a `Delete` outcome is already well over 99%.

Discussion Breakdowns

So late arrivers had high success rates, but did little to shift probabilities in the forecast model. I replicate that finding with users but with policy by singling out the `Snowball Clause` policy, summarized as:

Latest Citations	Avg. Rank	Forecast Shift Keep	Forecast Shift Delete	Earliest Citations	Avg. Rank	Forecast Shift Keep	Forecast Shift Delete
Civility	25.9	0.11	0.06	Living Person Biographies	2.7	0.31	0.12
No Personal Attacks	24.8	0.10	0.06	"Garage Bands"	3.3	N/A	0.09
Attack Pages	23.6	N/A	0.07	Notability (Media)	3.3	0.25	0.06
Disruptive Editing	22.6	0.12	0.04	Notability (Astronomy)	3.4	0.29	0.12
Gaming the System	21.1	0.09	0.06	Notability (Martial Arts)	4.0	0.27	0.09
Arguments to Avoid	20.4	0.13	0.08	Notability (Music)	4.1	0.21	0.08
Ignore All Rules	19.4	0.13	0.07	No Hoaxes	4.6	0.16	0.05

Table 9: Policies sorted by the ordinal rank of when they appear in discussion, and the mean forecast shift of votes where that citation appears, split by vote label. Many early-appearing policies overlap with the influential notability policies from Table 4.

"If an issue does not have a snowball's chance in hell of being accepted by a certain process, there's no need to run it through the entire process." This policy is cited once it is clear that consensus has been reached and that there is no need to hold discussion open for the full seven days. Indeed, votes citing this policy have the highest success rate and *lowest* forecast shift of any policy in the taxonomy. Citing the `Snowball` policy in `Keep` votes is similarly indicative of a contribution that will not change the forecast probabilities.

It is also possible to examine other policies that appear very late, with little forecast shift. Unlike `Snowball`, which almost always appears in successful votes, now let's focus on policy citations that appear late and are *not* successful. I sort policies by the mean ordinal rank of the post in which they appear; in Table 9, I present the top-ranked policies on each end of this measure (for clarity, referencing that table: `WP:Civility` citations appear in the 26th contribution to a discussion, on average). These policies are procedural and often indicate a breakdown in debate, with little information for the model to shift the likely outcome of debate. Instead, they are indicators that the debate's content-focused discussion has ended, an outcome is highly likely, and debate decorum has now broken down entirely. This includes citations to policies like `No personal attacks`, `Gaming the System`, and `No legal threats`. Voters that cite these policies are on the losing side of debates, posting very late, and also are not changing the direction of those debates in which they appear.

Eastern Michigan University student life [edit]

Eastern Michigan University student life (edit | talk | history | links | watch | logs | views) – (View log)

(*Find sources:* Google (books · news · newspapers · scholar · free images · WP refs) · FENS · JSTOR · NYT · TWL)

- This material belongs (and already is) in the main article for Eastern Michigan University. The topic isn't notable and extensive enough to warrant a separate article. ElKevbo (talk) 19:22, 31 May 2011 (UTC)
- **Keep**, at least for now. The material is in the main article because it was merged in, despite no consensus for a merge; one editor (Pwojdacz) was in favor and one editor (me) was opposed, but Pwojdacz went ahead and merged. As for whether the topic is extensive enough to warrant its own article, I think 11k of content (including 4,500+ characters of readable prose) suggests that it is. The main article at Eastern Michigan University is fairly large. Wikipedia:Article size suggests that "Readers may tire of reading a page much longer than about 30 to 50 KB, which roughly corresponds to 6,000 to 10,000 words of readable prose." Recent versions of that article range from 35 to 42 kB, comfortably within that range. Wikipedia:Summary style states that "Sections of long articles should be spun off into their own articles leaving a summary in its place" and that is basically what was done here. There is ongoing debate, both at the article talk page and at Talk:Eastern Michigan University about the future structure of subarticles dealing with the different aspects of EMU. Until that debate is settled, it seems premature to delete this article. cmadler (talk) 20:05, 31 May 2011 (UTC)

Figure 19: Citations in low-success rate votes that cause little change in forecasts come late in discussions, often citing detailed technical policies rather than focusing on persuasion or notability.

Previous work from Joyce et al.[203], studying a randomly selected set of 588 debates, focused in on that the `WP:IAR` policy has a significant effect on the present study's much larger corpus. That policy has been cited a total of 1,361 times in this corpus; among the votes in which the policy was cited, `Keep` votes were successful 10.3% less often than in the corpus overall, and appearing in successful `Keep` votes 53.7% of the time compared to a 64.0% baseline, and `Delete` votes dropped in success rates by 12.2%.

[203] Elisabeth Joyce, Jacqueline C Pike, and Brian S Butler. "Rules and roles vs. consensus: Self-governed deliberative mass collaboration bureaucracies". In: *American Behavioral Scientist* 57.5 (2013), pp. 576–594

Less contentious but also prevalent is procedural citation to editing guidelines, such as `Editing Policy` and `Article Size`. These votes, typically used in debates about lists that have been separated out of main articles and into separate standalone pages, tend to come very late in discussions. An example in Figure 19, arguing about whether a subset of information about a regional university merits its own article. In the example, I again omit some of the discussion, earlier votes that trended toward `Delete`. The late arrival of a `Keep` voter citing structural policy about preferable length of Wikipedia articles was basically futile; citation to policy on spinning out articles is made, but consensus has already been reached.

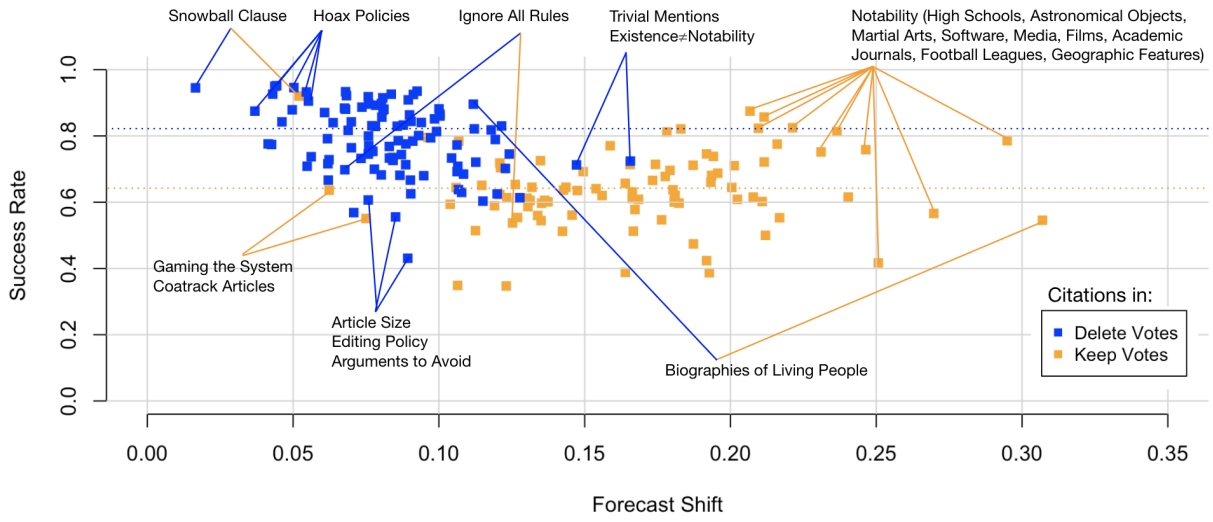


Figure 20: Summary of success rates and forecast shifts for various policies. Scatter plot shows all policy pages with at least 25 citations in either `Keep` or `Delete` votes. Dotted lines mark baseline success rates.

A scatter plot showing the full distribution of policies analyzed for this study appears in Figure 20. I separate policies by their appearance in `Delete` and `Keep` votes. Policies that I highlighted earlier in this analysis are labeled. This is a busy figure, so let's take the time below to analyze its component pieces in detail. Overall, I find that because `Delete` votes are more successful, so too are citations that appear in those votes, but that observing any one of these votes does not tend to produce a large shift in probable outcomes in the forecast model. As a result, policy citations from `Delete` votes clus-

ter in the top left of the scatter plot. Citations in `Keep` votes cause a much greater shift in the forecast model, as seen by the nearly clean partitioning of blue and orange clusters in the scatter plot. These policy citations are not necessarily successful, but do make the final outcome far less certain for the forecast model.

Future Directions

Tools for Decision Support

The potential for algorithmic decision-making as a support aid and tool in AfD is high. The use of machine learning tools powered by NLP has precedent in that community: tools *already* exist and are in widespread use for numerous behind-the-scenes tasks like vandalism detection[204], bot detection[205], and article quality assessment[206]. But it is not obvious which action is appropriate to take when administrator decisions disagree with predictions from forecasts, opening a broader question of trust in machine learning systems. Can this model be used to recognize when a poor decision is being made, or when participating editors are missing key experience levels or subject matter expertise? Future implementations of forecast models for Wikipedia may be able to notice maladaptive behaviors in groups, and recommend either a pause in decision-making when a "surprising" outcome is being chosen by an administrator, or could even be extended to active recruiting of new voices that are potentially underrepresented in existing discussion.

[204] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. "Building automated vandalism detection tools for Wikidata". In: *Proceedings of the International Conference on the World Wide Web*. 2017, pp. 1647–1654

[205] Andrew Hall, Loren Terveen, and Aaron Halfaker. "Bot Detection in Wikidata Using Behavioral and Other Informal Cues". In: *Proceedings of CSCW* (2018), p. 64

[206] Aaron Halfaker. "Interpolating quality dynamics in wikipedia and demonstrating the keilana effect". In: *Proceedings of WikiSym*. ACM. 2017, p. 19

Tempering any optimism about a technology-centered intervention, though, we must also consider the role of stakeholders and participants within the *AfD* process. Predicting the relative effectiveness of technological interventions is complicated; some approaches to improving retention have worked well[207], other attempts have backfired and been shown to *decrease* productivity of new users[208]. Any discussion of algorithmic interventions will need to be tempered by the unsteady reception to bots in the Wikipedia editorial system in general[209]. This work is not meant to propose technology-first solutions for the broader systemic and structural challenges with inequity in *AfD*, but to open a quantitative discussion of the existing norms and outcomes for marginalized users in this context. I hope it opens a fruitful avenue for future work to explore and a promising way to turn explanation into real-world action based not necessarily on bots or algorithms, but on clear-headed understanding of the data that describes the status quo and the actions that may effect change.

[207] Jonathan T Morgan et al. "Tea and sympathy: crafting positive new user experiences on wikipedia". In: *Proceedings of CSCW*. ACM. 2013, pp. 839–848

[208] Jodi Schneider, Bluma S Gelley, and Aaron Halfaker. "Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review". In: *Proceedings of the international symposium on open collaboration*. ACM. 2014, p. 26

[209] Richard Stuart Geiger II. "Robots.txt: An Ethnographic Investigation of Automated Software Agents in User-Generated Content Platforms". PhD thesis. University of California, Berkeley, 2015

Of course, future explanatory research will undoubtedly benefit from extension beyond just forecast shift, the primary tool I used in my citation analysis. This metric was just one way of recognizing

what is going on in discussions. It aligns neatly with findings from past work, and does not require any explicit labeling of preferences or votes at the granularity of turns or even individuals. This makes the metric well-suited to discussion contexts where no votes may be explicitly recorded. But in the *AfD* context, with the advantage of full discussion logs and explicit votes and outcomes, metrics that take more advantage of discussion structure may be appropriate. The highly structured hypergraph representation from Hua et al.[210] may serve as inspiration here. We can also use this model for deeper temporal analysis of Wikipedia's evolution over time, including a test of which policies have risen and fallen in prominence throughout the duration of the site's rise and decline.

[210] Yiqing Hua et al. "WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community". In: *Proceedings of EMNLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2818–2823. URL: http://aclweb.org/anthology/D18-1305

Influence on Group Decision-Making

There's a potentially more impactful future direction for this work – but impact is not always benevolent.

By focusing in on the social, we have a potential to build interventions for group debate. One next step for this research, and the measurement of forecast shift in debates generally, is to recognize, describe, and make visible the role of gatekeeping and identity-driven discourse behaviors in enforcing or even intensifying existing advantages for strategic long-term users. My existing work demonstrates that forecasts shift based on immediately observable characteristics, like how early a user is to arrive at a debate, how many posts they make, or how they cite policy. Interventions based on these features might measure and attempt to *alter* outcomes based on the factors that are known from our explanation of the domain.

Who would want to do this? For an optimistic view, consider the role of newcomers in debate. Over the last several years, Wikipedia has worked to bring newcomers into their community, but predicting what will be effective in such a complex domain. I want to know *who* gets a voice and a vote when arguing about what gets included in Wikipedia. Distinguishing the relative role of individuals will enable deeper process analysis of factors like diversity on teams[211], the interplay between individual participants and the process of resolving conflicts or disputes[212], and the granular habits that lead to effective outcomes.

These habits are often process-oriented, small-scale, and not adequately captured by survey or demographic variables[213], opening exciting new dimensions for behavioral science research. Specific, highly salient personal attributes include gender, as well as *tenure*, a user's prior experience based on their time since registration and initial participation in the community. A third factor of *prestige*, mea-

[211] Julia B Bear and Anita Williams Woolley. "The role of gender in team collaboration and performance". In: *Interdisciplinary science reviews* 36.2 (2011), pp. 146–153

[212] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. "Why differences make a difference: A field study of diversity, conflict and performance in workgroups". In: *Administrative science quarterly* 44.4 (1999), pp. 741–763

[213] Christoph Riedl and Anita Williams Woolley. "Teams vs. crowds: A field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-based problem solving performance". In: *Academy of Management Discoveries* 3.4 (2017), pp. 382–403

sured by administrator privileges or user profile awards, may also be relevant for study and intervention. These interventions based on identity may also be affected by the topic, context, and individuals participating in a specific discussion.

Future work may also benefit from measuring decision *quality*, a topic so far studied only by Lam et al.[214]. In that study, researchers identified poor decisions as those that were reversed at a later date: they flagged poor decisions either when an article was successfully re-nominated for deletion a previously kept article, or when a page that had previously been deleted as part of the *AfD* process was recreated. The magician from Figure 18, for instance, now has a recreated Wikipedia page with additional content. Another experiment would be to limit our analysis to close decisions with narrow margins, such as the 7.6% of votes ending in ties and the 5.2% of votes where administrators overruled the majority vote.

[214] Shyong K Lam et al. "WP: clubhouse?: an exploration of Wikipedia's gender imbalance". In: *Proceedings of WikiSym*. ACM. 2011, pp. 1–10

This focus on identity as a positive factor for group decision-making is not merely theoretical or aspirational. Lam et al. have already shown that diverse groups of decision-makers improves quality[215]. By going beyond raw statistics like edit count and into more granular, informed explanations of the strategies used in *AfD*, like policy citation, I have already shown there is nuance to be found in this domain. Next, by acknowledging the role of identity and diversity in group decision-making, I believe Wikipedia has an opportunity to greatly improve the quality of decision-making, breadth of representation, and level of participation in their community. This thesis leaves ample room for further study.

[215] Shyong K Lam et al. "WP: clubhouse?: an exploration of Wikipedia's gender imbalance". In: *Proceedings of WikiSym*. ACM. 2011, pp. 1–10

But let's also acknowledge the adversarial uses of machine learning and data analytics for influencing decision-making. Years ago, Facebook researchers manipulated News Feed contents to test their ability to influence emotions[216]; just like I experienced in the essay scoring domain, this produced a flurry of news controversy when the implications of the study dawned on the broader media[217]. This was just a shadow of the controversy to come shortly thereafter, though – the Cambridge Analytica scandal in the lead-up to the 2016 election[218]. The results showed what can be done with a data-driven approach to recognizing successful influences on decision-making and injecting algorithmic behavior into the mix.

[216] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8788–8790

[217] Robinson Meyer. "Everything We Know About Facebook's Secret Mood Manipulation Experiment". In: *The Atlantic* (2014). Accessed 2020-08-01. https://bit.ly/30l57kA

[218] Adrien Chen. "Cambridge Analytica and our lives inside the surveillance machine". In: *The New Yorker* 21 (2018), pp. 8–10

And so this work on Wikipedia can't be left on its own, looking only to beneficial uses. Late in this dissertation I'll engage in discussion on the potential for analysis of algorithmic tools evaluated not just in a vacuum, but in a *power* hierarchy, focusing on the funders of the work we do.

Part III: Automated Essay Scoring

Automated essay scoring (AES) mimics the judgment of educators evaluating the quality of student writing. There is a lot of reasonable concern, both practically and pedagogically, about what that mimicry means for schools, teachers, and students.

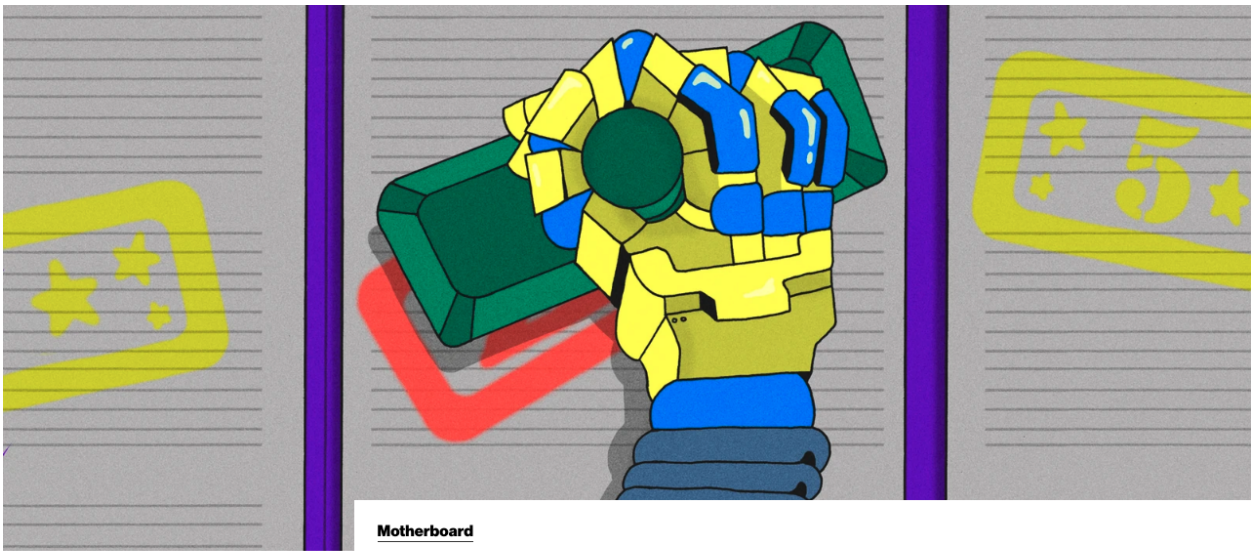
The work in this section is inspired by nearly a decade of work. After starting this section of the thesis with a brief overview of the historical context of AES, I report the results of an investigation on neural methods compared to classical machine learning, published at an ACL workshop this year[219].

The section continues with the specific project I spent the most time on during this thesis research, a partnership with the University at Albany on a series of evaluations of DAACS. The results of that study comprise two publications, in preparation for later this year:

- The first demonstrates the efficacy of the baseline machine learning models and investigates one particular explanatory phenomenon, the five-paragraph essay[220].
- The second investigates the potential for non-scoring-related topic models to explain student content choices, then digs in deeper on the specific phenomenon of non-adherent writing.

[219] Elijah Mayfield and Alan W Black. "Should you Fine-Tune BERT for Automated Essay Scoring?" In: *Proceedings of BEA*. 2020

[220] Elijah Mayfield et al. "Five-Paragraph Essays and Fair Automated Scoring in Online Higher Education". In: *Assessing Writing*. Under review



Motherboard

Flawed Algorithms Are Grading Millions of Students' Essays

Fooled by gibberish and highly susceptible to human bias, automated essay-scoring systems are being increasingly adopted, a Motherboard investigation has found

Figure 21: An August 2019 *Vice* report on automated essay scoring brought renewed attention to automated essay scoring, this time in the context of implementations for Common Core standardized testing.

Context and Background

Writing, though central to education, is labor intensive to grade. Teachers must balance giving students iterative practice and feedback with their own scarce time. So schools turn to automation.

Originally used for summative purposes in standardized testing[221], automated essay scoring (AES) systems are now frequently found in classrooms[222], typically enabled by training data scored on reliable rubrics to give consistent and clear goals for writers[223]. Essay scoring has been an intensely studied and debated field, the subject of Kaggle competitions[224] and mainstream media press[225], particularly for assessment purposes in high-stakes testing. A large body of research has shown that these systems have error rates equivalent to, or even slightly lower than, scores from classroom teachers or hired scorers in high-volume testing settings like the GRE and TOEFL[226,227]. But skepticism toward algorithmic scoring of essays continues for a variety of reasons. Many composition pedagogy experts have expressed concerns about its role in writing education[228,229], and some scholars have also disputed the validity of the scores themselves, noting the high correlation between scores in standardized testing and raw features like word count[230].

These are grave allegations and are worth taking seriously; and there is another level of concern exists not in deployment of systems but in their very design. Cultural bias creeps into AES through rubric writing and scoring of training data, unless extensive countermeasures are taken to maintain reliability across student backgrounds and varied response types[231]. 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. But of course, it is this focused domain and well-formulated set of categorical output labels that makes the task so enticing for researchers.

Before going through the steps of building, evaluating, and defending an AES system, this chapter attempts to make sense of the history of this debate and the broader educational context of writing assessment in which these systems are used.

[221] Jing Chen et al. "Building e-rater® Scoring Models Using Machine Learning Methods". In: *ETS Research Report Series* 2016.1 (2016), pp. 1–12

[222] Joshua Wilson and Rod D Roscoe. "Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy". In: *Journal of Educational Computing Research* (2019), p. 0735633119830764

[223] Y Malini Reddy and Heidi Andrade. "A review of rubric use in higher education". In: *Assessment & evaluation in higher education* 35.4 (2010), pp. 435–448

[224] Mark D Shermis and Ben Hamner. "Contrasting state-of-the-art automated scoring of essays: Analysis". In: *Proceedings of NCME*. 2012, pp. 14–16

[225] John Markoff. "Essay-Grading Software Offers Professors a Break". In: *The New York Times* (Apr. 2013). (Accessed on 06-30-2020.) URL: https://nyti.ms/2BoUaoF

[226] Yigal Attali and Jill Burstein. "Automated Essay Scoring with e-Rater® V. 2.0". In: *ETS Research Report Series* 2 (2004)

[227] Mark Shermis and Jill Burstein. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013

[228] NCTE. *NCTE Position Statement on Machine Scoring*. https://bit.ly/3dQHaVY. Accessed 2020-06-30. 2013. URL: https://bit.ly/3dQHaVY

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

[230] Les Perelman. "When "the state of the art" is counting words". In: *Assessing Writing* 21 (2014), pp. 104–111

[231] Anastassia Loukina et al. "Using exemplar responses for training and evaluating automated speech scoring systems". In: *Proceedings of BEA*. 2018, pp. 1–12

Overview of Writing Assessment

Students who learn to think of writing as a process that includes iterative improvement demonstrate large gains in transferable skills[232,233]. Unfortunately, this process is difficult to learn and complex to teach, needing differentiated instruction across students and incorporating strategies that may vary across tasks [234]. Teachers tend to view this element of instruction as difficult and time- consuming, and rarely teach the revision process in depth [235].

An open question is whether biases in assessment are more strongly embedded in the requirements of the assignments themselves, or the minds and preference of individual instructors doing the grading. On one hand, the bias influence of individual teachers is large. Biases of instructors are well established; grading is significantly influenced by mechanical errors like punctuation and capitalization when grading writing, even when using a content-specific rubric; but research has shown these personal preferences extend to student gender, physical attractiveness, and even penmanship[236]. On the other, the shaping of assignments and grading support like rubrics have a large effect on how students are measured. Teachers feel more confident when using rubrics, and the use of those rubrics increases inter-rater reliability in scoring[237,238]. Rubrics can be useful in higher education, with their effectiveness primarily tied to how well raters are trained and the appropriateness of the rubric design to the task being studied. Additionally, there is significant potential for formative use by students.

This reliance on rubrics has the potential impact, though, of narrowing the definition of good writing. This narrowing toward a limited range of essay structures pushes students away from culturally relevant writing styles and forms more representative of the writing they encounter in their day-to-day life[239]. For students of color, schools systematically devalue their home language use[240,241], creating a "double consciousness" where students use language one way at home and another way at school[242]. These expectations in practice grant extra privilege to affluent White students, both for the language they use[243] and their norms for behaviors like help-seeking[244].

[232] Stephanie Dix. ""What did I change and why did I do it?" Young writers' revision practices". In: *Literacy* 40.1 (2006), pp. 3–10

[233] Marion Tillema et al. "Relating self reports of writing behaviour and online task execution using a temporal model". In: *Metacognition and Learning* 6.3 (2011), pp. 229–253

[234] John Hayes and Linda Flower. "Identifying the Organization of Writing Processes". In: *Cognitive Processes in writing*. Ed. by L Gregg and E Teinber. Erlbaum, 1980

[235] Steve Graham and Karen Harris. "Writing Better: Effective Strategies for Teaching Students with Learning Difficulties." In: *Brookes Publishing Company* (2005)

[236] Ali Reza Rezaei and Michael Lovorn. "Reliability and validity of rubrics for assessment through writing". In: *Assessing writing* 15.1 (2010), pp. 18–39

[237] Heidi L Andrade, Ying Du, and Xiaolei Wang. "Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing". In: *Educational Measurement: Issues and Practice* 27.2 (2008), pp. 3–13

[238] Y Malini Reddy and Heidi Andrade. "A review of rubric use in higher education". In: *Assessment & evaluation in higher education* 35.4 (2010), pp. 435–448

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

[240] John R Rickford and Russell John Rickford. *Spoken soul: The story of black English*. Wiley New York, 2000

[241] David Holbrook. "Native American ELL Students, Indian English, and the Title III Formula Grant". In: *Annual Bilingual/Multicultural Education Conference*. 2011

[242] Anne H Charity Hudley and Christine Mallinson. *We Do Language: English Variation in the Secondary English Classroom*. Teachers College Press, 2013

[243] H Samy Alim and Geneva Smitherman. *Articulate while Black: Barack Obama, language, and race in the US*. Oxford University Press, 2012

[244] Jessica McCrory Calarco. ""I need help!" Social class and children's help-seeking in elementary school". In: *American Sociological Review* 76.6 (2011), pp. 862–882

Writing Assessment, Gender, and Race

Part of the reason that educators are skeptical of algorithms for scoring essays is that so many of the challenges students face in improving their writing scores are present at entrenched levels, far beyond what can be done with algorithmic interventions.

For instance, the culture of school writing is gendered. Gender is also a known predictor of educational outcomes: male students tend to underperform female students, and the gap tends to increase with age[245]. Many of these studies do not include transgender students, who are systematically under-supported and under-represented in educational research, either erased entirely or receiving limited support grouped with broader LGBT student needs[246]. In early (high school) years, girls' writing is more preferred because it's more descriptive and empathetic. In university writing, bold, assertive, self-confident writing is preferred which tilts the scales to men, especially those from historically advantaged backgrounds[247]. All of these results mean that student writing exists in a social context full of discriminatory beliefs about "proper" school language[248,249]. Among school-age children, gender differences for student scores on writing tasks exist: at adolescent ages, girls have an advantage in measured writing skill, while boys struggle with writing, particularly personal narratives[250]. The gender gap increases with age, peaking at the oldest ages tested in K-12 education literature, typically around age 21. Scores among girls are also more "bunched" around a mean than boys, who have higher variance; as a result, girls' scores benefit from a curriculum focus on regular coursework more than boys[251]. Language and race is an even more heated topic. Race is the single strongest predictor of educational outcomes in the United States[252]. Black and Native students underperform relative to their White and Asian peers, for reasons tied to systemic failures to support those populations. The same is true of Hispanic students and recent immigrants from Asia, Latin America, and the Caribbean[253].

[245] Caroline Scheiber et al. "Gender differences in achievement in a large, nationally representative sample of children and adolescents". In: *Psychology in the Schools* 52.4 (2015), pp. 335–348

[246] John P Dugan, Michelle L Kusel, and Dawn M Simounet. "Transgender college students: An exploratory study of perceptions, engagement, and educational outcomes". In: *Journal of College Student Development* 53.5 (2012), pp. 719–736

[247] Becky Francis et al. "University lecturers' perceptions of gender and undergraduate writing". In: *British Journal of Sociology of Education* 24.3 (2003), pp. 357–373

[248] Anne Curzan. "Teaching the politics of standard English". In: *Journal of English Linguistics* 30.4 (2002), pp. 339–352

[249] Lippi-Green Rosini et al. *English with an accent: Language, ideology, and discrimination in the United States*. Psychology Press, 1997

[250] Caroline Scheiber et al. "Gender differences in achievement in a large, nationally representative sample of children and adolescents". In: *Psychology in the Schools* 52.4 (2015), pp. 335–348

[251] Jannette Elwood. "Equity issues in performance assessment: The contribution of teacher-assessed coursework to gender-related differences in examination performance". In: *Educational Research and Evaluation* 5.4 (1999), pp. 321–344

[252] Gloria Ladson-Billings. "From the achievement gap to the education debt: Understanding achievement in US schools". In: *Educational researcher* 35.7 (2006), pp. 3–12

[253] Sita G Patel et al. "The achievement gap among newcomer immigrant adolescents: Life stressors hinder Latina/o academic success". In: *Journal of Latinos and Education* 15.2 (2016), pp. 121–133

Differences in gender also intersect with race – for instance, in transitions between schools, male performance declines more than for female students, and Black males see the largest declines of all[254]. Finally, we have ample evidence to suggest that non-native English speakers struggle with writing in English compared to native speaking peers. Schools are not well-equipped to teach those students[255], and sets students up to lose interest in developing writing skills[256].

Writing Assessment in Higher Education

My work with DAACS was situated in university settings, not in K-12. But many of these challenges follow those students into higher education, tied to preparedness from high school, and to economic factors like additional financial burden and lack of support[257,258]. But factors at that level are more complex and represent an increasingly nontraditional student population[259]. Income, job status, military service, age, and first-generation students may also have differential outcomes relative to their peers. However, it is not clear how such demographics would interact with writing ability, because of the limited research literature for adult learners. Expectations about writing differ between college students and K-12 students[260]; there is reason to believe many of the findings from middle and high schools will not transfer to our population, as the population of college-attending students represents significant selection bias[261]; the students that comprise the statistics pointing to an achievement gap in K-12 never make it to college of any kind, changing the population being studied. These pathways and drop-off points produce an entirely different socioeconomic strata in higher education[262], meaning we cannot be sure what results will transfer.

History of Automated Essay Scoring

In 1966 Ellis Page developed Project Essay Grade (PEG), kicking off 50 years of research into AES[263]. Methodology and computational power have advanced substantially, but the task has remained largely consistent. In traditional non-automated, large-scale essay scoring, two trained experts typically score each essay. A third expert resolves disagreements between them. The task of an AES system is to use scores from this traditional scoring process to train models that can score new essays as reliably as any individual rater.

Let's define the task of an automated essay scoring (or AES) system, because it is narrower than the overall job of a writing instructor. Student writing is scored following a rubric broken down into subscores ("traits"). These scores are almost always integer-valued,

[254] April Sutton et al. "Who Gets Ahead and Who Falls Behind During the Transition to High School? Academic Performance at the Intersection of Race/Ethnicity and Gender". In: *Social problems* 65.2 (2018), pp. 154–173

[255] Christina Ortmeier-Hooper and Kerry Anne Enright. *Mapping new territory: Toward an understanding of adolescent L2 writers and writing in US contexts*. 2011

[256] Bonny Norton Peirce. "Social identity, investment, and language learning". In: *TESOL quarterly* 29.1 (1995), pp. 9–31

[257] Ben Backes, Harry J Holzer, and Erin Dunlop Velez. "Is it worth it? Postsecondary education and labor market outcomes for the disadvantaged". In: *IZA Journal of Labor Policy* 4.1 (2015), p. 1

[258] Walter R Allen et al. "From Bakke to Fisher: African American Students in US Higher Education over Forty Years". In: *RSF: The Russell Sage Foundation Journal of the Social Sciences* 4.6 (2018), pp. 41–72

[259] Tressie McMillan Cottom. *Lower ed: The troubling rise of for-profit colleges in the new economy*. New Press, The, 2017

[260] Arthur N Applebee and Judith A Langer. "The state of writing instruction in America's schools: What existing data tell us". In: *Albany, NY: Center on English Learning and Achievement* (2006)

[261] Jennie E Brand and Yu Xie. "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education". In: *American sociological review* 75.2 (2010), pp. 273–302

[262] Sara Goldrick-Rab. "Following their every move: An investigation of social-class differences in college pathways". In: *Sociology of Education* 79.1 (2006), pp. 67–79

[263] Ellis B Page. "The imminence of... grading essays by computer". In: *The Phi Delta Kappan* 47.5 (1966), pp. 238–243

usually with fewer than 10 possible score points. In most contexts, students respond to "prompts," a specific writing task with a limited range of potential answers, often tied to a specific source document or material to write about or analyze. Because of the narrowed vocabulary, consistent length, and relatively homogenous substance of these essays, copious prior research has shown that rubric-based scoring of prompt-specific writing can be scored reliably on a common rubric by trained annotators, and that this scoring can be replicated reliably by machine learning methods[264].

[264] Mark D Shermis. "State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration". In: *Assessing Writing* 20 (2014), pp. 53–76

It has typically been the guidance of researchers in the field, when working in collaboration with practitioners, to recommend relatively small training set sizes. In typical cases, automated essay scoring systems are trained on hundreds of essays; in high-stakes tests like the GRE or TOEFL, between 1,000 and 5,000 essays might be used for supervised learning. These numbers are large relative to a high school educator's classroom size, making personalization to individual instructors infeasible; but relative to many tasks in natural language processing, the numbers are positively tiny. Neural approaches to question answering, reading comprehension, and sentiment analysis are routinely trained on corpora consisting of hundreds of thousands or even millions of texts.

Figure 22: An example of rubric traits designed for use in automated essay scoring, from my previous work on Turnitin *Revision Assistant*.

Automated essay scoring has been built up out of the field of psychometrics, moreso than learning science or machine learning. That field focuses on building tools that use machine learning to mimic the judgment of educators evaluating the quality of student writing.

Originally used for summative purposes in standardized testing and the GRE[265], these systems are typically enabled by rubrics, which give consistent and clear goals for writers[266]. Student essays are scored either on a single holistic scale, or analytically following a rubric that breaks out subscores based on "traits" (as in Figure 22). These scores are almost always integer-valued, and typically have fewer than 10 possible score points, though scales with as many as 60 points exist. In most contexts, students respond to "prompts," a specific writing activity with predefined content, and only receive feedback on valid attempts to respond to the prompt. Applications typically include a "library" of many prompts that students can be assigned, at instructor discretion.

[265] Jing Chen et al. "Building e-rater® Scoring Models Using Machine Learning Methods". In: *ETS Research Report Series* 2016.1 (2016), pp. 1–12

[266] Y Malini Reddy and Heidi Andrade. "A review of rubric use in higher education". In: *Assessment & evaluation in higher education* 35.4 (2010), pp. 435–448

AES has focused historically on replicating expert readers for large-scale scoring of thousands of essays, either for end-of-year standardized assessments or entrance exams like the GRE or TOEFL[267]. While those industry vendors that provide these products have their own datasets, academic AES research has been dominated for the last five years by the dataset from the 2012 Automated Student Assessment Prize (ASAP) competition[268]. This competition used essays written to eight prompts, scored on a variety of scales. The competition had a private phase with companies as competitors, followed by a public Kaggle competition with anonymized data. Shermis[269] provides a summary of the competition, and most recent research papers report their results using the same public dataset[270,271].

[267] Yigal Attali and Jill Burstein. "Automated Essay Scoring with e-Rater® V. 2.0". In: *ETS Research Report Series* 2 (2004)

[268] `https://www.kaggle.com/c/asap-aes`

[269] Mark D Shermis. "State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration". In: *Assessing Writing* 20 (2014), pp. 53–76

[270] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. "Flexible domain adaptation for automated essay scoring using correlated linear regression". In: *Proceedings of EMNLP*. 2015, pp. 431–439

[271] Kaveh Taghipour and Hwee Tou Ng. "A neural approach to automated essay scoring". In: *Proceedings of EMNLP*. 2016, pp. 1882–1891

This use preferences interpretable model features informed by psychometrics, often representing high-level characteristics of writing like coherence or lexical sophistication. The primary goal is defensibility of the underlying model, known as construct validity. This construct validity through feature choice has been emphasized over measuring the ability to provide actionable guidance to writers based on the scoring. Let's look at how that might be changing.

Feedback

Instruction is still (hopefully!) at the center of education, and so direct instructional technologies using algorithmic decision-making is a core part of this thesis. Throughout the history of AES research there has been a recognition that formative feedback is an essential goal for automated tools to improve student learning[272,273]. Automated scoring brings value to the classroom, but targeted formative feedback alongside those scores is vital to the development of writing proficiency. Research in writing education demonstrates that localized, actionable feedback, presented as part of an iterative writing process, is effective. By connecting comments to the rubric's evaluation crite-

[272] Semire Dikli. "An overview of automated scoring of essays". In: *The Journal of Technology, Learning and Assessment* 5.1 (2006)

[273] Rod D Roscoe and Danielle S McNamara. "Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom." In: *Journal of Educational Psychology* 105.4 (2013), p. 1010

ria, students can use the feedback to foster their ability to reflect and self-assess. Self-reported responses to teacher feedback confirm[274] that students value a combination of positive and critical comments that are specific to their own writing, and connected to the evaluation criteria. Such feedback is especially valuable on preliminary drafts, instead of later in the writing process[275]. We also see new techniques aimed at improving AES having a ripple effect of advancing fields like argument mining[276] and rhetorical structure detection[277]. For writers who are proficient or already working in professional settings, language technologies provide scaffolds like grammatical error detection and correction[278].

In the 1990s and early 2000s, classroom technology was released based on this approach, including ETS Criterion, Pearson WriteToLearn, and Vantage MyAccess. Classroom reviews of these products were mixed at best. While their use positively impacted student writing[279], students felt negative about the experience[280]. Teachers using earlier tools stated that automated scoring must be paired with actionable next steps for writers[281]. Building on this, academic work has used AES to provide formative writing instruction and feedback that students perceive as "informative, valuable, and enjoyable"[282], providing more efficient learning gains than practice alone[283].

Alongside the emergence of that research, a newer generation of tools has refocused AES to prioritize feedback to students. These include TenMarks Writing, WriteLab, Grammarly, PEG Writing, and Turnitin Revision Assistant. AES feedback's impact on writing quality varies by product. For instance, PEG Writing has been shown to save teachers time and let them focus on higher-level writing skills, but not to improve writing quality[284]. To date, there is little work discussing the longitudinal effect of AES on classroom instruction during the school year.

In many cases, automated feedback is not directly driven by the scoring algorithms themselves. For instance, ETS Criterion uses the scoring models from its e-rater system, but provides the student with feedback based on a series of separate algorithms that detect usage and mechanics errors, particular aspects of style (e.g. passive voice), and detection of discourse elements. Arizona State's Writing Pal is an intelligent tutoring system that scaffolds writing and feedback within learning tasks. Its feedback focuses on things like structure and relevance, though it uses engineered essay features for each feedback type, divorced from the scoring model itself.

The exact approach toward providing students with feedback vary greatly across systems, but immediacy for student viewing and reduction of teacher workload are almost universally the primary goals for such classroom systems. Broadly speaking, teachers have

[274] Melanie R. Weaver. "Do students value feedback? Student perceptions of tutors' written responses". In: *Assessment & Evaluation in Higher Education* 31.3 (2006), pp. 379–394

[275] Dana R. Ferris. "Student Reactions to Teacher Response in Multiple-Draft Composition Classrooms". In: *TESOL Quarterly* 29.1 (1995), pp. 33–53

[276] Huy V Nguyen and Diane J Litman. "Argument mining for improving the automated scoring of persuasive essays". In: *AAAI Conference on Artificial Intelligence*. 2018

[277] James Fiacco, Elena Cotos, and Carolyn Rosé. "Towards Enabling Feedback on Rhetorical Structure with Neural Sequence Models". In: *Proceedings of LAK*. ACM. 2019, pp. 310–319

[278] Hwee Tou Ng et al. "The CoNLL-2014 shared task on grammatical error correction". In: *Proceedings of CONLL*. 2014, pp. 1–14

[279] Mark D Shermis, Cynthia Wilson Garvan, and Yanbo Diao. "The Impact of Automated Essay Scoring on Writing Outcomes". In: *Annual Meeting of the National Council on Measurement in Education (NCME)*. 2008

[280] Cassandra Scharber, Sara Dexter, and Eric Riedel. "Students' Experiences with an Automated Essay Scorer." In: *Journal of Technology, Learning, and Assessment* 7.1 (2008), n1

[281] Eric Riedel et al. "Experimental evidence on the effectiveness of automated essay scoring in teacher education cases". In: *Journal of Educational Computing Research* 35.3 (2006), pp. 267–287

[282] Rod D Roscoe and Danielle S McNamara. "Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom." In: *Journal of Educational Psychology* 105.4 (2013), p. 1010

[283] Scott Crossley et al. "Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system". In: *Proceedings of AIED*. Springer. 2013

[284] Joshua Wilson and Amanda Czik. "Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality". In: *Computers & Education* 100 (2016), pp. 94–109

been supportive of this technology as deployed in schools today. These applications follow the same algorithmic approaches as high-stakes scoring, in most cases; the key difference is in the use of the tool for practice rather than measurement.

Efficacy of AES for Learning

Up until recently, the only major study of a deployed system was from Grimes & Warschauer, studying Vantage *MyAccess* in a high school setting. In that work, teachers expressed significant reservations and viewed the AES models as "fallible," stating that automated scoring must be paired with feedback that gave meaningful next steps for writers[285]; nevertheless, the teachers expressed optimism for future development.

While more limited in scope, automated writing evaluation algorithms have seen some more limited use in higher education. Cotos et al., for instance, have shown that students can use feedback based on AES to better structure their introductions to writing in scientific genres[286]. Johnson et al. showed that students had nuanced understandings of the feedback they received such systems, and were able to engage with the revision process during writing with an AES support[287]. But overall, composition scholars have resisted the introduction of automated tools into their classrooms, instead recommending local, contextual solutions that are much more expensive to manage, but that produce more fine-grained and useful information about student ability[288].

All of the inertia in real-world use is moving AES technology away from use exclusively for saving money in high-volume and high-stakes psychometric assessments; instead, the future appears to be a formative experience for AES, focusing on feedback and student agency and growth, and lower-stakes recommendations in contexts that also include better infrastructure for valuing pre-existing faculty expertise, institutional support for writing, and professional development.

Why Explain AES?

The AES field grew out of a need to replicate the work of expert essay readers at minimal cost as the scale of assessments grew and expenses expanded past what was financially feasible. In this pursuit, many approaches try to directly incorporate insights from those experts that were being replaced. Many systems, including Ellis's PEG, ETS's e-rater[289], and more, focus on feature engineering. They create a small to moderate number of expert-designed features meant to

[285] Douglas Grimes and Mark Warschauer. "Utility in a fallible tool: A multi-site case study of automated writing evaluation." In: *Journal of Technology, Learning, and Assessment* 8.6 (2010)

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

[287] Adam C Johnson, Joshua Wilson, and Rod D Roscoe. "College student perceptions of writing errors, text quality, and author characteristics". In: *Assessing Writing* 34 (2017), pp. 72–87

[288] William Condon. "Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?" In: *Assessing Writing* 18.1 (2013), pp. 100–108

[289] Martin Chodorow and Jill Burstein. "Beyond essay length: evaluating e-rater®'s performance on toefl® essays". In: *ETS Research Report Series* 2004.1 (2004), pp. i–38

represent high level characteristics of writing[290]. These may include measures such as coherence or lexical sophistication[291]. The connection between the constructs used by human experts and the AES system is generally emphasized as a central feature.

An increasingly influential body of work attempts to avoid laborious feature engineering by using large numbers of low-level textual features[292] or neural network derived word or paragraph embeddings[293,294]. These systems use high dimensional modeling techniques, and relax the constraint that model features should mimic human reasoning. We use this approach, demonstrating with our feedback system that expert derived features are not required for interpretable output. Our results in this area are parallel to recent work in the deep learning domain on creating textual rationales for network predictions [295].

AES is potentially a powerful tool for supporting student writing, but that the effect was at least partially — and perhaps wholly — mediated by a broader cultural shift that supported teachers in their instructional role, rather than being an independent "silver bullet" that produced results on its own. One serious flaw in the way AES — and in some ways, technologies developed by learning science in general — has been defended is its narrow reliance on construct validity as a path to pedagogical defensibility. Arguments have focused on the expert judgment in feature engineering of AES models. In 2004, a defense of then-leading automated scoring model, ETS e-Rater, argued of its 12 features:

> *"[they] reflect essential characteristics in essay writing and are aligned with human scoring criteria [...] Validity here refers to the degree to which the system actually does what is intended, in this case, measuring the quality of writing."*[296]

.

This approach, while not without its flaws, has nevertheless been taken more seriously and led to use of these systems in high-profile standardized exams. But it does not speak to the primary worry of skeptical educators: not whether the system is reliably performing the job of standardized testing, but whether students are being treated with equity in mind throughout the education process, and whether the system furthers the goals of those students in their academic journey. These are hard questions, to be sure, but they aren't *impossible* questions. The only reason they appear intractable to psychometricians is their disinterest in the kinds of justifications those scientists are accustomed to making. I am interested in changing that approach.

Traditional models used in psychometrics for standardized tests rely on construct validity as a defense of their application to high-

[290] Danielle S McNamara et al. "A hierarchical classification approach to automated essay scoring". In: *Assessing Writing* 23 (2015), pp. 35–59

[291] Torsten Zesch and Oren Melamud. "Automatic generation of challenging distractors using context-sensitive inference rules". In: *Proceedings of BEA*. 2014, pp. 143–148

[292] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. "Flexible domain adaptation for automated essay scoring using correlated linear regression". In: *Proceedings of EMNLP*. 2015, pp. 431–439

[293] Kaveh Taghipour and Hwee Tou Ng. "A neural approach to automated essay scoring". In: *Proceedings of EMNLP*. 2016, pp. 1882–1891

[294] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. "Automatic Text Scoring Using Neural Networks". In: *Proceedings of ACL*. 2016, pp. 715–725

[295] Tao Lei, Regina Barzilay, and Tommi Jaakkola. "Rationalizing Neural Predictions". In: *Proceedings of EMNLP*. 2016, pp. 107–117

[296] Jill Burstein, Martin Chodorow, and Claudia Leacock. "Automated essay evaluation: The Criterion online writing service". In: *AI Magazine* 25.3 (2004), p. 27

stakes decision-making like university admissions; as a result, automated models are routinely deployed with fewer than 100 features, each of which is carefully tuned to align to an intuitive principle of a text's quality. More modern approaches, meanwhile, use standard NLP methods like *n*-grams or word embeddings to represent text using a less interpretable feature space.

Implications of a Defensible AES

An outstanding and controversial question in the AES literature is exactly what is learned by the AES classifiers on these small datasets, and what implications that has for instructional pedagogy and test prep. Are the models learning overly superficial features of writing, forcing students into a narrow task of replicating an idealized, simplistic essay form like the five-paragraph essay? Or are AES models capable of accurately evaluating and giving reliable scores even to more nontraditional texts that eschew the structure taught by tutors?

The answer to this question has ramifications for educational equity. Access to tutors and test prep for students is not distributed evenly. Affluent families from well-resourced suburbs and western countries have enormous advantages in educational attainment already, and an enormous body of work in education is based on closing "the achievement gap" produced by this inequity. On the other hand, much of that work comes in the form of intensive test-prep courses for students from marginalized backgrounds, which "teaches to the test" by having students memorize specific structural elements of essay texts, potentially at the expense of creativity and individual expression.

Developers of AES software have an opportunity for social change here. As technologists driving the policy conversation around the future use of algorithmic tools, we have enormous leverage in defining the tasks and training data that will feed into machine learning systems; we are listened to in a way that many other stakeholders are not[297,298]. But in order to do so, we need to understand whether our approaches work, and if so, why.

[297] David Lehr and Paul Ohm. "Playing with the Data: What Legal Scholars Should Learn About Machine Learning". In: *UCDL Rev.* 51 (2017), p. 653

[298] Kenneth Holstein et al. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of CHI*. 2018

Evaluating Neural Methods

Deep neural networks dominate today's NLP research. In particular, publishing research in NLP today almost requires interacting with the Transformer architecture popularized by BERT[299]. These models use large volumes of existing text data to pre-train multilayer neural networks with context-sensitive meaning of, and relations between, words. The models, which often consist of over 100 million parameters, are then fine-tuned to a specific new labeled dataset and used for classification.

[299] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL*. 2019

Automation of writing assessment has historically relied on simpler models, like multivariate regression from a small set of justifiable variables chosen by psychometricians[300]. This produces models that retain direct mappings between variables and recognizable characteristics of writing, like coherence or lexical sophistication[301,302]. In psychometrics more generally, they have a term for this defense of a model – "construct validity" – built on rigorously defined alignment of model features to recognizable skills[303].

[300] Yigal Attali and Jill Burstein. "Automated Essay Scoring with e-Rater® V. 2.0". In: *ETS Research Report Series* 2 (2004)

[301] Helen Yannakoudakis and Ted Briscoe. "Modeling coherence in ESOL learner texts". In: *Proceedings of BEA*. Association for Computational Linguistics. 2012, pp. 33–43

[302] Sowmya Vajjala. "Automated assessment of non-native learner essays: Investigating the role of linguistic features". In: *International Journal of Artificial Intelligence in Education* 28.1 (2018), pp. 79–105

[303] Yigal Attali. "Validity and Reliability of Automated Essay Scoring". In: *Handbook of automated essay evaluation: Current applications and new directions* (2013), p. 181

This thesis has already laid out how perilous it is to work with deep neural models for causal explanation of machine learning predictions. Add on the need for construct validity, which is outsizedly important in AES, and acknowledging the results from the first part of this thesis on causal explanation[304,305], and you get to an incredible tight spot for model defensibility. As I set forth to build a defensible, explainable machine learning system for AES in this dissertation, one core question and the main focus of this chapter is whether a move to Transformers is worth the cost.

[304] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of NAACL*. 2019

[305] Sofia Serrano and Noah A Smith. "Is Attention Interpretable?" In: *Proceedings of ACL*. 2019

The chief technical contribution of this chapter is to measure the results of using BERT, when fine-tuned, for AES tasks. I describe an experimental setup with multiple levels of technical difficulty from bag-of-words models to fine-tuned Transformers, and show that the approaches perform similarly. While Transformers do match state-of-the-art accuracy, they do so at the expense of hardware constraints, including an up to 100x slowdown in training time. My data shows that Transformer models improve on n-gram baselines by no more than 5%. Training a full Transformer architecture requires major hardware and energy expenditure during training, increases the carbon footprint of machine learning[306], and opens questions of fairness and explanation that are extremely hard to answer; as I'll

[306] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". in: *Proceedings of ACL* (2019)

discuss later in the dissertation, they are also hard to prioritize in organizational settings[307].

Relative to the rest of this thesis, this chapter is narrowly focused. My point is this: in AES, the payoff for the extra effort of a fine-tuned Transformer is minimal, as human inter-rater reliability often creates a ceiling for model performance. I intend to set the stage for my work on DAACS with methods that make use of neural methods only when helpful, and only with the knowledge that a better, alternate road to explanation and defensibility will be necessary. But I don't want to leave any reader with the impression that deep learning has *no* potential to improve performance far beyond current baseline levels from classical methods; and so I conclude the chapter with discussion of some of the directions where we might see rapid gains from that more state-of-the-art approach, even if they are not prioritized in the rest of my own work here.

[307] Michael A Madaio et al. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI". in: *Proceedings of CHI*. 2020, pp. 1–14

Background

Natural language processing has historically used *n*-gram bag-of-words features to predict labels for documents. These were the standard representation of text data for decades and are still in widespread use [308]. More recently, the field moved over to word *embeddings*, where words are represented not as a single feature but as dense vectors learned from large unsupervised corpora. Early approaches to dense representations using latent semantic analysis have been a major part of the literature on AES[309,310], but these were corpus-specific representations; more recent work was general-purpose and allowed for broader similarities between words to be captured, resulting in popular off-the-shelf representations like GloVe[311]. This allows similar words to have approximately similar representations, effectively managing feature sparsity.

But the greatest recent innovation has been *contextual* word embeddings, based on deep neural networks and in particular, Transformers. Rather than encoding a word's semantics as a static vector, these models adjust the representation of words based on their context in new documents. With multiple layers and sophisticated use of *attention mechanisms*[312], these newer models have outperformed the state-of-the-art on numerous tasks, and are currently the most accurate machine learning models on a very wide range of tasks[313,314]. The most popular architecture, BERT, produces a 768-dimensional final embedding based on a network with over 100 million total parameters in 12 layers.

These neural models are just starting to be used in machine learning for AES. This is especially the case for research identifying con-

[308] Dan Jurafsky and James H Martin. *Speech and language processing*. Vol. 3. Pearson London, 2014

[309] Peter W Foltz, Sara Gilliam, and Scott Kendall. "Supporting content-based feedback in on-line writing evaluation with LSA". in: *Interactive Learning Environments* 8.2 (2000), pp. 111–127

[310] Tristan Miller. "Essay assessment with latent semantic analysis". In: *Journal of Educational Computing Research* 29.4 (2003), pp. 495–512

[311] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of EMNLP*. 2014, pp. 1532–1543

[312] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *Proceedings of ICLR*. 2015

[313] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of NeurIPS*. 2017, pp. 5998–6008

[314] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *Proceedings of ACL*. 2019

structs as an intermediate representation for automated essay scoring and feedback[315,316]. But end-to-end models, where texts are the only input and scores are learned directly, are in their infancy in AES and have been used only in exploratory studies in the past year, particularly in work by Rodriguez et al.[317].

BERT for Automated Essay Scoring

For document classification, BERT is "fine-tuned" by adding a final layer at the end of the Transformer architecture, with one output neuron per class label. When learning from a new set of labeled training data, BERT evaluates the training set multiple times, each termed an *epoch*. A loss function, propagating backward to the model parameters, allows the model to learn relationships between the class labels in the new data and the contextual meaning of the words in the text. A learning rate determines the amount of change to a model's parameters. Extensive results have shown that careful control of the learning rate in a *curriculum* can produce an effective fine-tuning process[318]. While remarkably effective on many tasks, the NLP community is only just beginning to identify exactly what is *learned* in this process; research in "BERT-ology" is ongoing[319,320,321].

To date, there are no best practices on fine-tuning Transformers for AES; in this section I present options. I begin with classical machine learning, starting with traditional bag-of-words approaches and non-contextual word embeddings, used with Naïve Bayes and logistic regression classifiers, respectively. I then describe multiple curriculum learning options for fine-tuning BERT using AES data and best practices on fine-tuning. I end with two approaches based on BERT but with reduced hardware requirements.

Bag-of-Words Representations

The simplest and most longstanding features for document classification tasks is to represent documents as a "bag-of-words." In the simplest version of this task, surface *n*-grams of length 1-2 are extracted and given "one-hot" binary values indicating presence in a document. In prior AES results, an extension of this baseline to *n*-grams based on part-of-speech tags (of length 2-3), to capture syntax independent of content, and character *n*-grams of length 3-4, to provide robustness to misspellings, further improves AES performance[322,323]. This high-dimensional representation typically has a cutoff threshold where rare tokens are excluded: in this implementation, all *n*-grams without at least 5 token occurrences in training data. Even still, this is a sparse feature space with thousands of dimensions.

[315] James Fiacco, Elena Cotos, and Carolyn Rosé. "Towards Enabling Feedback on Rhetorical Structure with Neural Sequence Models". In: *Proceedings of LAK*. ACM. 2019, pp. 310–319

[316] Farah Nadeem et al. "Automated Essay Scoring with Discourse-Aware Neural Models". In: *Proceedings of BEA*. 2019, pp. 484–493

[317] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. "Language models and Automated Essay Scoring". In: *arXiv preprint arXiv:1909.09482* (2019)

[318] Leslie N Smith. "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay". In: *arXiv preprint arXiv:1803.09820* (2018)

[319] Olga Kovaleva et al. "Revealing the Dark Secrets of BERT". in: *Proceedings of EMNLP*. vol. 1. 2019, pp. 2465–2475

[320] Ganesh Jawahar, Benoıt Sagot, and Djamé Seddah. "What Does BERT Learn about the Structure of Language?" In: *Proceedings of ACL*. 2019, pp. 3651–3657

[321] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "Bert rediscovers the classical nlp pipeline". In: *Proceedings of ACL*. 2019

[322] Bronwyn Woods et al. "Formative essay feedback using predictive scoring models". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 2071–2080

[323] Brian Riordan, Michael Flor, and Robert Pugh. "How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models". In: *Proceedings of BEA*. 2019, pp. 116–126

For learning from bag-of-words representations, I use a Naïve Bayes classifier with Laplace smoothing, as implemented in Scikit-learn[324], with part-of-speech tagging from SpaCy[325].

[324] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12 (2011), pp. 2825–2830

[325] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* (2017)

Word Embeddings

A more modern representation of text uses word-level embeddings. This produces a dense vector, typically of up to 300 dimensions, representing each word in a document. In this implementation, I represent each document as the term-frequency-weighted mean of word-level embedding vectors from GloVe[326]. Unlike one-hot bag-of-words embeddings, this representation has real-valued features and Naïve Bayes models are inappropriate, so I instead train a logistic regression classifier, with the LibLinear solver[327] and L_2 regularization, as implemented in Scikit-learn.

[326] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of EMNLP*. 2014, pp. 1532–1543

[327] Rong-En Fan et al. "LIBLINEAR: A library for large linear classification". In: *Journal of machine learning research* 9.Aug (2008), pp. 1871–1874

Fine-Tuning BERT

In this work, I fine-tune BERT using the Fast.ai library. I selected this library because of its visibility to first-time users of deep learning and accessible online learning materials[328]. For new practitioners, their default settings are likely the most straightforward route to using deep learning.

[328] https://course.fast.ai/

Fast.ai recomends use of *cyclical* learning rate curricula for fine-tuning. In this policy, an upper and lower bound on learning rates are established. lr_{max} is a hyperparameter defining the maximum learning rate in one epoch of learning. I set $lr_{max} = 0.00001$. A lower bound is then derived from the upper bound, $lr_{min} = 0.04 * lr_{max}$.

In cyclical learning, the learning rate for fine-tuning begins at the lower bound, rises to the upper bound, then descends back to the lower bound. A high learning rate midway through training acts as regularization, allowing the model to avoid overfitting and avoiding local optima. Lower learning rates at the beginning and end of cycles allow for optimization within a local optimum, giving the model an opportunity to discover fine-grained new information again.

Here I assess three different curricula for cyclical learning rates, visualized in Figure 23. In the default approach, a maximum learning rate is set and cycles are repeated until reaching a threshold; for this work's halting criterion, I directly measure validation set accuracy of the model. Because of noise in deep learning training, halting at *any* decrease can lead to premature stopping; it is preferable to allow some small drop in performance at times. My implementation halts when accuracy on a validation set, measured in quadratic weighted kappa, decreases by over 0.01. In the second, "two-rate" approach[329], I follow the same algorithm, but at the first halting epoch, I instead

[329] Leslie N Smith. "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay". In: *arXiv preprint arXiv:1803.09820* (2018)

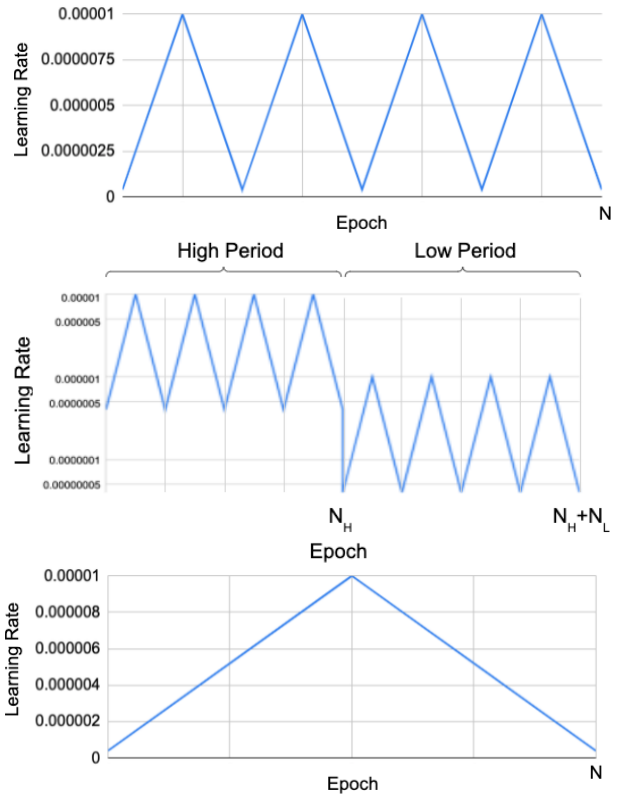


Figure 23: Illustration of cyclical (top), two-period cyclical (middle, log y-scale), and 1-cycle (bottom) learning rate curricula over N epochs.

backtrack by one epoch to a previously saved version of the network, then restart the training with a learning rate of $1 * 10^{-6}$ (one order of magnitude smaller). Finally, in the "1-cycle" policy, training is condensed into a single rise-and-fall pattern, spread over N epochs. Defining the exact training time N is a hyperparameter tuned on validation data. Finally, while BERT is optimized for sentence encoding, it is able to process documents up to 512 words long. I truncate a small number of essays longer than this maximum, mostly in ASAP dataset #2, where essays were much longer than in other datasets.

Feature Extraction from BERT

Fine-tuning is computationally expensive and can only run on GPU-enabled devices. Many practitioners in low-resource settings may not have access to appropriate cloud computing environments for these techniques. Previous work has described a compromise approach for using Transformer models without fine-tuning. In Peters et al.[330], the authors describe a new pipeline. Document texts are processed with an untuned BERT model; the final activations from network on the `[CLS]` token are then used directly as contextual word embeddings. This produces a 768-dimensional feature vector, representing

[330] Matthew E Peters, Sebastian Ruder, and Noah A Smith. "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP-2019)*. 2019, pp. 7–14

the full document. The values of this representation are then used as inputs for a traditional linear classifier (in thiscase, a logistic regression). In the context of education technology, a similar approach was described in [331] as a baseline for comparison in evaluating language-learner essays. This process allows us to use the world knowledge embedded in BERT without requiring fine-tuning of the model itself, and without need for GPUs at training or prediction time. For training, I use a logistic regression classifier as described in the section on GloVe.

[331] Farah Nadeem et al. "Automated Essay Scoring with Discourse-Aware Neural Models". In: *Proceedings of BEA*. 2019, pp. 484–493

DistilBERT

I am not the first researchers to question the value of full-scale Transformer models. Particularly in response to the carbon concerns of [332] and the desire for Transformer-based prediction on-device without access to cloud compute, [333] introduce DistilBERT, which they argue is equivalent to BERT in most practical aspects while reducing parameter size by 40% to 66 million, and decreasing model inference time by 60%. This is accomplished using a distillation method [334] in which a new, smaller "student" network is trained to reproduce the behavior of a pretrained "teacher" network. Once the smaller model is pretrained, interacting with it for the purposes of fine-tuning is identical to interacting with BERT directly. In this work, I only present results for DistilBERT with the "1-cycle" learning rate policy.

[332] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". in: *Proceedings of ACL* (2019)

[333] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019)

[334] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network". In: *stat* 1050 (2015), p. 9

Experiments

To test the overall impact of fine-tuning in the AES domain, I use five English-language datasets from the ASAP competition, jointly hosted by the Hewlett Foundation and Kaggle.com. This set of essay prompts was the subject of intense public attention and scrutiny in 2012 and its public release has shaped the discourse on AES ever since[335]. I discard the three datasets - prompts 1, 7, and 8 - with a scale of 10 or more possible points. This is not typical for most AES contexts. Prompts 2-6 are scored on smaller rubric scales with 4-6 points, a much more common scenario. I use the original, deanonymized data from[336]; an anonymized version of these datasets is available through Kaggle.com for public reproduction of results[337]. In all cases, human inter-rater reliability (IRR) is an approximate upper bound on performance. Reliability above human IRR is possible, as all models are trained on *resolved* scores that represent two scores plus a resolution process for disagreements between annotators.

[335] Mark D Shermis. "State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration". In: *Assessing Writing* 20 (2014), pp. 53–76

[336] Mark D Shermis and Ben Hamner. "Contrasting state-of-the-art automated scoring of essays: Analysis". In: *Proceedings of NCME*. 2012, pp. 14–16

[337] https://www.kaggle.com/c/asap-aes

Metrics and Baselines

For measuring reliability of automated assessments, I use a variant of Cohen's κ, with quadratic weights for "near-miss" predictions on an ordinal scale (QWK). This metric is standard in the AES community[338]. High-stakes testing organizations differ on exact cutoffs for acceptable performance, but threshold values between 0.6 and 0.8 QWK are typically used as a floor for testing purposes; human reliability below this threshold is generally not fit for summative student assessment.

In addition to measuring reliability, I also measure training and prediction time, in seconds. As this work seeks to evaluate the practical tradeoffs of the move to deep neural methods, this is an important secondary metric. For all experiments, training was performed on Google Colab Pro cloud servers with 32 GB of RAM and an NVideo Tesla P100 GPGPU.

I compare the results of BERT against several previously published benchmarks and results.

- Human IRR as initially reported in the Hewlett Foundation study[339].
- Industry best performance, as reported by eight commercial vendors and one open-source research team, also from that initial release of the Hewlett Foundation study.
- An early deep learning approach using a combination CNN+LSTM architecture that outperformed most reported results at that time[340].
- Two recent results using traditional non-neural models: our own from my time at Turnitin[341], which uses *n*-gram features in an ordinal logistic regression, and Cozma et al.[342], which uses a mix of string kernels and word2vec embeddings in a support vector regression.
- Rodriguez et al.[343], the one previously-published work that uses BERT as an AES classifier, along with comparisons to the similar XLNet architecture[344].

Experimental Setup

Following past publications, I evaluate all datasets using 5-fold cross-validation. Each of the five datasets contains approximately 1,800 essays, resulting in folds of 360 essays each. Additionally, for measuring loss when fine-tuning BERT, I hold out an additional 20% of

[338] David M Williamson, Xiaoming Xi, and F Jay Breyer. "A framework for evaluation and use of automated scoring". In: *Educational measurement: issues and practice* 31.1 (2012), pp. 2–13

[339] Mark D Shermis. "State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration". In: *Assessing Writing* 20 (2014), pp. 53–76

[340] Kaveh Taghipour and Hwee Tou Ng. "A neural approach to automated essay scoring". In: *Proceedings of EMNLP*. 2016, pp. 1882–1891

[341] Bronwyn Woods et al. "Formative essay feedback using predictive scoring models". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 2071–2080

[342] Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. "Automated essay scoring with string kernels and word embeddings". In: *Proceedings of ACL*. 2018

[343] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. "Language models and Automated Essay Scoring". In: *arXiv preprint arXiv:1909.09482* (2019)

[344] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Proceedings of NeurIPS*. 2019, pp. 5753–5763

each training fold as a validation set, meaning that each fold has approximately 1,150 essays used for training and 300 essays used for validation. I report mean QWK across the five folds. For measurement of training and prediction time, I report the sum of training time across all five folds and all datasets. For slow-running feature extraction, like *n*-gram part-of-speech features and word embedding-based features, I tag each sentence in the dataset only once and cache the results, rather than re-tagging each sentence on each fold. Finally, for models where distinguishing extraction from training time is meaningful, I present those times separately.

Results

Accuracy Evaluation

My primary results are presented in Table 10. I find, broadly, that all approaches to machine learning replicate human-level IRR as measured by QWK. Nearly eight years after the publication of the original study, no published results have exceeded vendor performance on three of the five prompt datasets; in all cases, a naive *n*-gram approach underperforms the state-of-the-art in industry and academia by 0.03-0.06 QWK. Fine-tuning with BERT also reaches approximately this performance, slightly underperforming previous results.

Model	**2**	**3**	**4**	**5**	**6**
Human IRR	.80	.77	.85	.74	.74
Hewlett	**.74**	**.75**	.82	**.83**	.78
Taghipour	.69	.69	.81	.81	.82
Woods	.71	.71	.81	.82	**.83**
Cozma	.73	.68	**.83**	**.83**	**.83**
Rodriguez	.70	.72	.82	.82	.82
n-grams	.71	.71	.78	.80	.79
Embeddings	.42	.41	.60	.49	.36
BERT-CLR	.66	.70	.80	.80	.79
BERT-1CYC	.64	.71	.82	.81	.79
BERT Features	.61	.59	.75	.75	.74
DistilBERT	.65	.70	.82	.81	.79
n-gram Gap	-.05	.00	.04	.01	.00

Table 10: Performance on each of ASAP datasets 2-6, in QWK, and execution time, in seconds. The final row shows the gap in QWK between the best-performing neural model and the *n*-gram baseline.

Of particular note is the low performance of GloVe embeddings relative to either neural or *n*-gram representations. This is surprising: while word embeddings are less popular now than deep neural methods, they still perform well on a wide range of tasks[345]. Few publications have noted this negative result for GloVe in the AES

[345] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of ACL*. 2014, pp. 238–247

domain; only Dong[346] uses GloVe as the primary representation of ASAP texts in an LSTM model, reporting lower QWK results than any baseline I presented here. One simple explanation for this may be that individual keywords matter a great deal for model performance. It is well established that vocabulary-based approaches are effective in AES tasks[347] and the lack of access to specific word-based features may hinder semantic vector representation. Indeed, only one competitive recent paper on AES uses non-contextual word vectors: Cozma et al.[348]. In this implementation, they do use word2vec, but rather than use word embeddings directly they first cluster words into a set of 500 "embedding clusters." Words that appear in texts are then counted in the feature vector as the centroid of that cluster - in effect, creating a 500-dimensional bag-of-words model.

Rodriguez et al.[349] demonstrate that it is possible to improve the performance of BERT slightly through hyperparameter optimization and a full grid search of possible settings. Sophisticated approaches like gradual unfreezing, discriminative fine-tuning, or increased counts of parameters through newer deep learning models consistently produces slight upticks in performance. I do not claim these results are the best that could be achieved with BERT fine-tuning, but instead argue that the ceiling of results at inter-rater reliability makes the optimization questionable.

[346] Fei Dong, Yue Zhang, and Jie Yang. "Attention-based recurrent convolutional neural network for automatic essay scoring". In: *Proceedings of CONLL*. 2017, pp. 153–162

[347] Derrick Higgins et al. "Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring". In: *arXiv preprint arXiv:1403.0801* (2014)

[348] Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. "Automated essay scoring with string kernels and word embeddings". In: *Proceedings of ACL*. 2018

[349] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. "Language models and Automated Essay Scoring". In: *arXiv preprint arXiv:1909.09482* (2019)

Runtime Evaluation

My secondary evaluation of models is based on training time and resource usage; those results are reported in Table 11. Here, deep learning approaches on GPU-enabled cloud compute produce an approximately 30-100 fold increase in end-to-end training time compared to a naive approach. In fact, this understates the gap, as approximately 75% of feature extraction and model training time in the naive approach is due to part-of-speech tagging rather than learning. Using BERT features as inputs to a linear classifier is an interesting compromise option, producing slightly lower performance on these datasets but with only a 2x slowdown at training time, all in feature extraction, and potentially retaining some of the semantic knowledge of the full BERT model. Further investigation should test whether additional features for intermediate layers, as explored in Peters et al.[350], is merited for AES.

Figure 24 explores this gap in training runtime more closely. Essays in the prompt 2 dataset are longer persuasive essays and are on average 378 words long, while datasets 3-6 correspond to shorter, source-based content knowledge prompts and are on average 98-152 words long. The need for truncation in dataset #2 for BERT, but

[350] Matthew E Peters, Sebastian Ruder, and Noah A Smith. "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP-2019)*. 2019, pp. 7–14

Model	F	T	P	Total
Embeddings	93	6	1	100
n-grams	82	27	2	111
BERT Features	213	10	1	224
DistilBERT	1,972		108	2,080
BERT-1CYC	2,956		192	3,148
BERT-CLR	11,309		210	11,519

Table 11: Cumulative experiment runtime, in seconds, of feature extraction (F), model training (T), and predicting on test sets (P), for ASAP datasets 2-6 with 5-fold cross-validation. Models with 1-cycle fine-tuning are measured at 5 epochs.

not for other approaches, may explain the underperformance of the model in that dataset. Additionally, differences across datasets highlight two key differences for fine-tuning a BERT model:

- Training time increases linearly with number of epochs and with average document length. As seen in Figure 24, this leads to a longer training for the longer essays of dataset #2, nearly as long as the other datasets combined.
- Performance converges on human inter-rater reliability more quickly for short content-based prompts, and performance begins to decrease due to overfitting in as few as 4 epochs. By comparison, in the longer, persuasive arguments of dataset 2, very small performance gains on held-out data continued even at the end of these experiments.

Figure 24 also presents results for DistilBERT. This work verifies prior published claims of speed improvements both in fine-tuning and at prediction time, relative to the baseline BERT model: training time was reduced by 33% and prediction time was reduced by 44%. This still represents at least a 20x increase in runtime relative to *n*-gram baselines both for training and prediction.

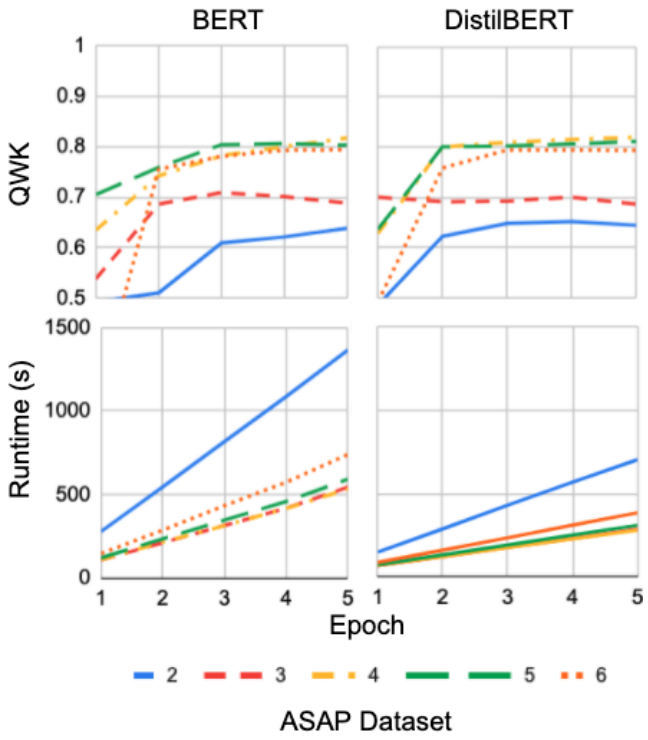


Figure 24: QWK (top) and training time (bottom, in seconds) and for 5-fold cross-validation of 1-cycle neural fine-tuning on ASAP datasets 2-6, for BERT (left) and DistilBERT (right).

Discussion on Neural Methods

The results in this chapter suggests that for scoring prompt-specific AES with reliable training sets, all approaches described – with the exception of non-contextual word embeddings – produce similar reliability, at approximately identical levels to human inter-rater reliability. There is a substantial increase in technical overhead required to implement Transformers and fine-tune them to reach this performance, with questionable practical gain compared to simpler baselines. The policy lesson for NLP researchers is that using deep learning for scoring alone is unlikely to be justifiable, given the slowdowns at both training and inference time, and the additional hardware requirements. In AES, at least, Transformer architectures are a hammer in search of a nail.

As I move to work on DAACS data specifically, BERT does come up and is used for feature extraction. But it is used in conjunction with traditional methods, as my data here shows that they cannot be ruled out as effective and simple methods for good performance. With DAACS data in particular, the small total training set sizes and relatively low reliability compared to industry datasets results will be a good fit for those methods. So let's get started.

Training and Auditing DAACS

I now focus in on a partnership for studying these problems with real-world data. My study investigates an open source tool, DAACS, that has been used at two private, non-profit online universities since 2017 to provide automated online support for incoming students. Up until 2019, these were the only sites where the tool had been used; these universities focus on non-traditional students, mid-career professionals, and military veterans. This system is currently in use by thousands of students annually. Later, in 2019, the tool was extended to use at a traditional brick-and-mortar school, as I'll investigate near the end of this chapter.

There are several tasks to investigate in this first section on DAACS, but much like in the Wikipedia analysis from earlier, the fundamental goal is to demonstrate that the machine learning system I am building is accurate enough to reliably make predictions that are worth explaining. Is the system making predictions that are reliable and informative enough to be worth the time to build a good explanation?

Alongside evaluation of overall human inter-rater reliability and automation reliability, I also conduct a *fairness* audit of the DAACS classifier. This approach to evaluating bias is now standard in the NLP literature: we'll measure disparate outcomes for different user groups based on demographics. Since the ProPublica investigative journalism on COMPAS in 2016, this has now become a standard way to measure models for equitable use[351]. That analysis, plus the explanatory work that follows, sets up a conversation about how an automated writing assessment system can be defensible and fair.

[351] Julia Angwin et al. "Machine bias". In: *ProPublica, May* 23 (2016)

Setting Description

DAACS was developed in partnership between the University at Albany and Excelsior College as part of a FIPSE "First in the World" grant. Students use DAACS in a multi-stage process. Students take a series of four assessments, including assessments for mathematics, reading, and self-regulated learning (SRL), and writing. The assessments provide feedback and guidance for students directly, and provide online dashboards to academic advisors, with summaries of students' individualized needs and available resources.

When used by students and integrated into existing advising structures at participating universities, initial results show that

DAACS is associated with students finishing more credits and staying in school in their first year[352]. These results are, at minimum, a predictive tool to recognize risk in first-time students, and are potentially causally linked to student long-term persistence, directly improving support for students.

[352] Jason Bryer and Angela M. Lui. "Efficacy of the Diagnostic Assessment and Achievement of College Students on Multiple Success Indicators". In: *Proceedings of AERA* (2019)

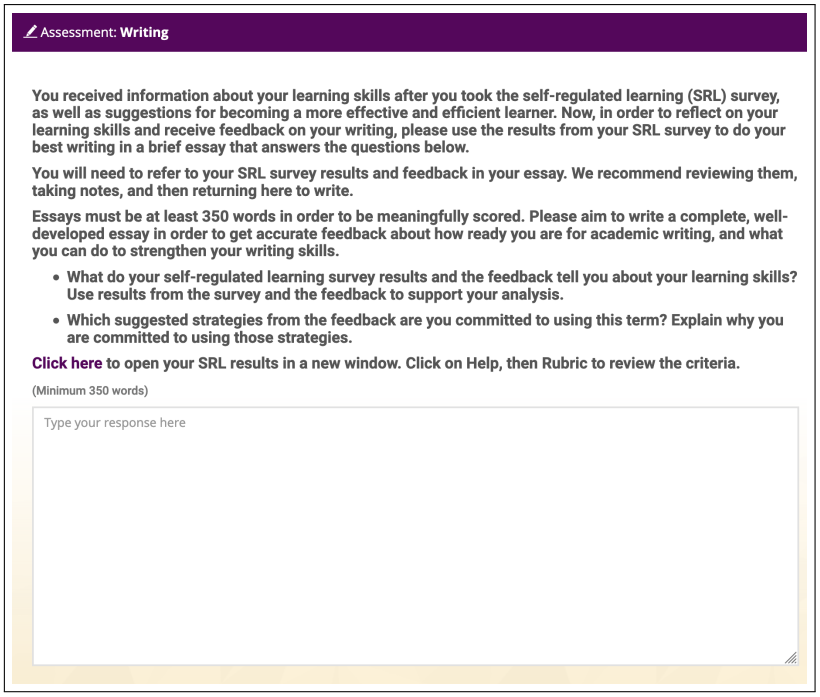


Figure 25: Screenshot from DAACS including the writing prompt students responded to for this dataset.

Within this context, the writing prompt seen in Figure 25 is available immediately after students complete their self-regulated learning assessment. It asks students to compose a brief essay of at least 350 words, in which they reflect on their SRL survey results and select and commit to using the strategies recommended for managing and improving their learning strategies, metacognition, and academic motivation. The dataset consists of 6,243 English-language essays submitted to this prompt from within the DAACS platform in a live implementation, collected between April 2017 and February 2018. In addition to essay text, I also have access to demographic information for students, including age, race, and gender, which will come into play later. Using those scored essays, I train AES on that rubric, and establish the baseline ability of machine learning to reproduce human judgment. I show which technical approaches work most effectively at reproducing different rubric traits, like high-level content traits, paragraph-level organization traits, and low-level sentence complexity and conventions traits.

Trait	Subtrait	Developing (1)	Emerging (2)	Mastering (3)
Content	Summary	The discussion of the survey and feedback is vague, poorly grounded in the survey results and feedback, and/or simplistic.	The essay uses evidence from survey results and feedback to summarize student's strengths and weaknesses in terms of self- regulated learning. The summary lacks sufficient detail; might be under-developed in places, e.g., strengths or weaknesses might get short shrift.	The essay uses relevant survey results and feedback to provide a detailed summary of both the student's strengths and weaknesses in terms of self-regulated learning.
Content	Suggestions	Choices of suggestions to which to commit are vague, if present at all, and/or only loosely connected to the survey results and feedback,if at all. The essay might refer to the continued use of current strategies but not to anything new related to the SRL feedback.	Choices of suggestions to which to commit are discussed. The connections to the survey and feedback are present but might not always be explicit.	The discussion of suggestions for improvement in SRL are logically and explicitly related to the survey results and feedback, and developed in sufficient depth.
Organization	Structure	The structure and order of the essay is weak, unclear, and/or illogical.	The essay has a general structure and order but may not have a clear overall organization that enables a reader to follow the progression of one idea to another. Although the structure is logical, it might seem haphazard at times. Note: One-sentence paragraphs do not necessarily reflect a problem with organization, but numerous such paragraphs might signal a weak or haphazard structure.	The essay is well-organized, with an order and structure that present the discussion in a clear, logical manner.

Table 12: Rubric used for scoring the 2017 DAACS data (part 1).

Trait	Subtrait	Developing (1)	Emerging (2)	Mastering (3)
Organization	Transitions	Transitions between paragraphs are missing or ineffective; paragraphs tend to abruptly shift from one idea to the next. Note: One-paragraph essays receive a 1 for this criterion.	Paragraphs are usually linked with transitions, as needed. The transitions might be implied or strained, but the reader can follow along.	Transitions between paragraphs are appropriate and effective, and strengthen the progression of the essay (e.g. "The second aspect . . ." "The last aspect . . ." and/or the repetition of important ideas and terms to connect paragraphs).
Paragraphs	Focus	Most or all paragraphs lack one clear, main point; might have several topics. Note: Numerous brief paragraphs of one or two sentences each might indicate a problem with paragraph focus and warrant a score of 1.	Paragraphs are generally but not consistently focused on a main idea or point. Some paragraphs might lack a clear focus in an essay in which the majority of paragraphs maintain a clear focus on a main idea.	Paragraphs are consistently and clearly focused on a main idea or point.
Paragraphs	Cohesion	The connections between ideas in sentences within paragraphs are unclear. Little effective use of linking words and phrases.	The ideas or information in each sentence within a paragraph are generally but not consistenly linked together, if only loosely. Additional or better choices of linking words and phrases would clarify the connections b/w ideas within paragraphs.	Within paragraphs, the individual sentences are seamlessly linked together; the reader can see the relationship between the ideas or information in one sentence and those in another sentence. The writing explicitly links sentences and ideas using adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which), conjunctions (e.g., and, or, while, whereas), and/or the repetition of key words, as appropriate.

Table 13: Rubric used for scoring the 2017 DAACS data (part 2).

Trait	Subtrait	Developing (1)	Emerging (2)	Mastering (3)
Sentences	Correct	Significant syntax problems, such as fragments, run-on sentences, missing/extra words, awkward constructions, dangling modifiers, and/or transposed words, are present and numerous enough to distract readers and impede meaning.	Grammatically incorrect sentences, when present, are minor and do not interfere with meaning.	There are very few or no significant syntax problems. The writer is capable of managing even complex syntactic structures correctly.
Sentences	Complex	The sentences lack syntactic complexity and vary little, if at all, in structure. The sentences tend to be relatively simple in structure, following a basic subject-verb-object pattern perhaps with a few additional elements, such as brief introductory phrases, prepositional phrases, or modifiers.	Complex syntactic structures are present but may not always be managed effectively; sentence structures may be varied but are not often sophisticated.	Consistent and appropriate use of a variety of sentence structures, including sophisticated sentence structures, such as complex, compound, or compound-complex sentences, and other complex syntactic forms, such as extended participial phrases and relative clauses.
Conventions		A pattern of errors in spelling, punctuation, usage (such as incorrect word forms or subject-verb agreement), and/or capitalization suggest that the writer struggles with the rules for conventions.	Spelling, punctuation, usage, and capitalization are generally correct. There may be errors but there is no pattern that suggests that the writer struggles with the basic rules.	Spelling, punctuation, and capitalization are correct to the extent that almost no editing is needed. There are very few, if any, very minor errors of usage.

Table 14: Rubric used for scoring the 2017 DAACS data (part 3).

Hand-Scoring

Rubric

The DAACS rubric has five broad categories. First, two content dimensions measure the quality of the ideas present in a text: the clarity of the student's `SUMMARY` [1] of their self-regulated learning assessment results, which are presented to them in terms of metacognition, learning strategies, and motivation sections; and `SUGGESTIONS` [2], measuring how well they describe a forward-looking plan based on those results. Next, four traits assess larger-scale document-level measures of writing quality: the `STRUCTURE` [3] of the text's organization, the `TRANSITIONS` [4] of the document from one paragraph to another, the `COHESION` [5] of the sentence within paragraphs, and then the `FOCUS ON A MAIN IDEA` [6] with a clear focus wiithin each paragraph. Finally, three traits measure students' sentence-level composition. Sentences are evaluated for their `COMPLEXITY` [7] and their grammatical `CORRECTNESS` [8]. The final trait measures student ability to use the `CONVENTIONS` [9] of written English spelling, capitalization, and punctuation.

This instrument was developed with a number of validity goals, and the criteria for each score point were written in collaboration with the Director of an undergraduate writing program, who is also the director of a local site of the National Writing Project. The full rubric as used in scoring the 2017 data is available in Table 12-14.

Inter-Rater Reliability

Scoring for a training dataset was completed on a subset of 540 essays. This data was double-scored following industry norms and achieved acceptable levels of inter-rater reliability, and the scoring was completed prior to my involvement in the project. Statistics are shown in Table 15. Four metrics are given: exact accuracy, adjacent accuracy (which only counts disagreements as errors if they are off by more than 1 point), as well as Cohen's Kappa with both linear and quadratic weights. The last of these is the preferred metric in the automated essay scoring literature, accounting for chance agreement and giving "partial credit" for close disagreements.

Raters perform with moderate reliability on most high-level content and structural features. For lower-level scoring of sentence-level traits, usage, and conventions, raters achieve only fair inter-rater reliability. The observed level of reliability would not be suitable for high-stakes testing, where industry vendors recommend QWK of at least 0.70 [353]. However, as a low-stakes diagnostic tool, the reliability was deemed sufficient for preliminary use with student data.

[353] David M Williamson, Xiaoming Xi, and F Jay Breyer. "A framework for evaluation and use of automated scoring". In: *Educational measurement: issues and practice* 31.1 (2012), pp. 2–13

Table 15 also presents the correlation between word count and score on each trait. Perelman[354] found that the correlation on timed writing samples from the well-known Hewlett Foundation study ranged from 0.434 to 0.785; on average, almost half of all variance in scores was explained by word count alone. In our data, this pattern is not observed. Instead, the highest correlation between word count and sentence complexity, at 0.199, ranging all the way down to a very small inverse correlation between word count and adherence to grammatical conventions, at -0.037. Across all traits there is low correlation between word count and score.

[354] Les Perelman. "When "the state of the art" is counting words". In: *Assessing Writing* 21 (2014), pp. 104–111

		Exact	Exact+Adjacent	κ	QWK	Length Correlation
Content	Summary	58.0	94.9	0.325	0.485	0.081
	Suggestions	58.6	94.8	0.341	0.547	0.072
Organization	Structure	62.9	99.0	0.282	0.409	0.164
	Transitions	57.8	96.9	0.343	0.526	0.197
Paragraph	Focus	61.9	96.3	0.304	0.522	0.136
	Cohesion	59.7	98.0	0.227	0.338	0.133
Sentence	Correctness	55.2	96.6	0.240	0.403	0.050
	Complexity	58.6	98.8	0.237	0.338	0.199
Conventions		55.2	98.8	0.248	0.403	-0.037

Table 15: Inter-rater reliability based on human judgment.

Automated Scoring Methods

Machine learning for automated essay scoring relies on a meaningful representation of text as quantifiable, low-level features. These features are then aggregated to identify statistical patterns that are most common at each score point on a rubric. The previous chapter showed that blindly making use of the state-of-the-art is *not* an ideal option, and that simpler approaches can often perform similarly or even better; so for DAACS in particular I leaned on those results to craft several more straightforward options to optimize within.

Based on what we now know about what machine learning methods work in this domain, to complete this work I implemented three different machine learning paradigms, each of which is commonly known in the machine learning literature. First I use a basic set of features relying on superficial text characteristics, which require only minimal automated processing of texts without any content extraction. Second, I extract features using classical "bag-of-words" natural language processing. Finally, I implement a feature extraction technique using deep neural networks, which are the current state-of-the-art in the broader field of natural language processing.

Surface Methods: In the most simplistic approach to automated essay scoring, content is ignored completely. Instead, straightforward, easily calculated metrics like word count and average sentence length are made directly available to a logistic regression classifier. While composition scholars have historically chafed at the use of methods like these, they serve as a useful minimum baseline precisely because many writing rubrics, past and present, correlate closely with these surface-level features.

For this method, I make four count features available to all machine learning models:

- The number of total characters in the text.
- The number of total words, separated by whitespace.
- The number of total paragraphs, separated by line breaks.
- A binary feature with a value of 1 when an essay contains at least one paragraph break.

Note that I allow both the classical and neural models below to have access to these surface features, in addition to their much more sophisticated content-based representation. This is based on copious evidence that word count is a factor in essay scoring, and prior published research, including my own and the results earlier in this thesis, showing that explicitly modeling length as a feature actually

isolates and reduces machine learning's dependency on length as a feature for scoring.

Classical Methods: The most common historical representation, called "bag-of-words," was the standard representation of text data for decades and is still in widespread use today. As my previous chapter showed, this approach is still more than viable in the context of AES. This approach, used by most commercial vendors, lists hundreds or thousands of very simplistic binary features are extracted, each representing a single vocabulary word or phrase of length 1, 2, or 3 (unigrams, bigrams, and trigrams; collectively, *n*-grams). The value of each feature is set to 1 only if the word or phrase appears in the text. These representations are "sparse" - they consist of an enormous number of relatively rare features, and most features have a value of 0 for most essays. A Naïve Bayes classifier is most commonly used to then make a prediction about an essay's score based on how many of those features were observed in a given text.

In this classical implementation, I use a categorical Naïve Bayes classifier with surface *n*-grams of length 1 and 2, as well as *n*-grams based on part-of-speech tags of length 2 and 3. Part-of-speech tags are well-known in the automated assessment literature as a useful set of proxies for syntactic structure in student writing, allowing models to abstract away from the specific vocabulary used by students. This implementation closely follows these best practices.

Neural Methods: Over the last decade in the state-of-the-art in natural language processing, best practices have shifted to contextual word embeddings, based on deep neural networks. Rather than encoding an essay as a collection of observed vocabulary words, these models generate large-scale numeric vectors that represent text in context. Additionally, rather than learning solely on the basis of a single training set, they make use of enormous amounts of background data scraped off of websites, novels, and other sources of language. As a result, they are able to "embed" pre-existing world knowledge of the meaning of words and phrases, quantifying how similar or different those words and phrases are. This makes the models more resilient to identifying patterns based on synonyms, implicit relationships between words, and stylistic tendencies, while making the models less reliant on matching exact keywords for automated classification.

The most popular architecture, the Transformer, is based on a network with hundreds of millions of total parameters in a series of a dozen or more interconnected layers; state-of-the-art models are hundreds of megabytes in size, pre-trained by large tech companies like Google and Facebook, and released for open source use. While the models are still opaque to interpretability[355], these models are

[355] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A Primer in BERTology: What we know about how BERT works". In: *arXiv preprint arXiv:2002.12327* (2020)

currently the most accurate machine learning models on a very wide range of tasks[356,357] and are used in all major technology fields.

Nevertheless, adoption of deep neural models in automated essay scoring is nascent. Early approaches to dense representations using latent semantic analysis have been a major part of the literature on AES[358,359], but these were dataset-specific representations without pre-training. Zhang & Litman, for instance, demonstrated the feasibility of neural methods for evidence extraction[360], but determined the models were too difficult to interpret at this time for practical use in their followup study working directly with students[361].

Part of the reason for the relative weakness of neural methods here, as I showed in the previous chapter, may be the relative simplicity of the task; as shown in past studies, classical machine learning techniques are sufficient for reaching human levels of reliability on rubric scoring of content traits; automation typically cannot produce more reliable scoring than the innate subjectivity of underlying human judgment. Additionally, the use of pre-training adds even more concerns around the "black box" nature of automated scoring, as the neural models have been shown to learn and even amplify the stereotypes and biases that are found in the data used for pre-training[362,363]. Finally, using these neural methods requires substantially more sophisticated hardware including cloud servers and GPU processing, while taking an order of magnitude more processing power and time to train and make predictions.

For the implementation of deep learning in this work, I extract dense contextual representations from the most popular Transformer neural model, BERT[364]. The representation of each essay is extracted from the final output layer for each sentence in each essay. A sentence's encoding from the pretrained BERT model is a vector with 768 real-valued dimensions, representing the contextual meaning of that sentence as a point in a high-dimensional space. I average these sentence-level encodings, weighted by the natural logarithm of the number of tokens in each sentence, resulting in one shared vector representation of the entire essay. This parallels recommendations from the natural language processing literature for feature extraction from neural networks on relatively straightforward tasks with small datasets[365]. Evidence from the previous chapter has already shown that this stripped-down version of BERT may have some value but that the full-scale BERT implementation is likely not worth the extensive extra technical effort and maintenance cost.

[356] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of NeurIPS*. 2017, pp. 5998–6008

[357] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *Proceedings of ACL*. 2019

[358] Peter W Foltz, Sara Gilliam, and Scott Kendall. "Supporting content-based feedback in on-line writing evaluation with LSA". in: *Interactive Learning Environments* 8.2 (2000), pp. 111–127

[359] Tristan Miller. "Essay assessment with latent semantic analysis". In: *Journal of Educational Computing Research* 29.4 (2003), pp. 495–512

[360] Justine Zhang et al. "Characterizing online public discussions through patterns of participant interactions". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–27

[361] Elaine Lin Wang et al. "eRevis (ing): Students' revision of text evidence use in an automated writing evaluation system". In: *Assessing Writing* (2020), p. 100449

[362] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Proceedings of NeurIPS*. ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 4349–4357

[363] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of NAACL*. 2019, pp. 609–614

[364] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL*. 2019

[365] Matthew E Peters, Sebastian Ruder, and Noah A Smith. "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP-2019)*. 2019, pp. 7–14

		Human	Surface	Classical	Neural	QWK Gap
Content	Summary	0.485	0.097	0.472	0.372	-0.013
	Suggestions	0.547	0.091	0.505	0.463	-0.042
Organization	Structure	0.409	0.220	0.265	0.287	-0.122
	Transitions	0.526	0.247	0.234	0.333	-0.193
Paragraph	Focus	0.522	0.559	0.160	0.547	+0.037
	Cohesion	0.338	0.235	0.216	0.319	-0.019
Sentence	Correctness	0.403	0.022	0.149	0.377	-0.026
	Complexity	0.338	0.137	0.247	0.366	+0.028
Conventions		0.403	0.049	0.167	0.336	-0.067

Table 16: Comparison of automated essay scoring to human baseline inter-rater reliability (in QWK), and resulting gap between humans and the best-performing automated method.

Automated Scoring Reliability

Automated essay scoring cannot be measured for inter-rater reliability in quite the same way as reliability between two human raters. Machine learning that includes an essay in training data has overfit to that essay – the quality of the model cannot be evaluated on that essay, because it has already "seen" the correct answers. Instead, the standard methodology is to use cross-validation.

As seen in Table 16, tuned scoring with machine learning performs close to human reliability but with significant variability in the representations that are most effective. For content-based scoring, the classical methods used in traditional NLP systems outperforms newer neural models, and comes very close to human inter-rater reliability. For all other traits with one exception, newer neural methods that have the pre-training advantage of real-world language use outperform those older techniques. This matches the observation from prior work that, while classical methods struggle on grammatical features of writing, newer deep learning approaches are “unreasonably” effective[366].

There are a few traits where performance does not fit this overall pattern. On the document-level organization traits, while neural models still outperform older methods, they significantly underperform human inter-rater reliability. The most concerning statistic comes on the “Focus” trait, simple surface features not only outperform other automated methods, they also outperform the inter-rater reliability between two trained humans. Neural methods, which do take content into account, are also able to match human inter-rater reliability.

One consistent pattern of note is that automation across most traits has a specific tendency to under-report the minority class, whichever label appears least frequently in the data. Across the DAACS dataset,

[366] Dimitris Alikaniotis and Vipul Raheja. “The Unreasonable Effectiveness of Transformer Language Models in Grammatical Error Correction”. In: *Proceedings of BEA*. 2019, pp. 127–133

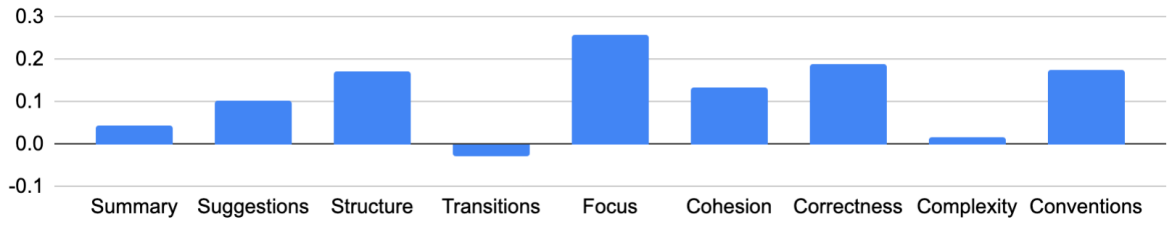


Figure 26: Shift in population mean scores when using AES, compared to hand-scoring.

this is always the lowest score (1) on each trait. As a result, when moving to automation, the population mean is inflated slightly on all but one trait, as shown in Figure 26.

Williamson et al. recommends that automated scoring degrade inter-rater reliability by no more than 0.10 QWK relative to human inter-rater reliability[367]. My results shows that the best-performing model in all cases other than document-level organizational traits approximately match human performance; automated performance on organizational traits with these models would not be ready for replacing human judgment, even for summative-only scoring.

[367] David M Williamson, Xiaoming Xi, and F Jay Breyer. "A framework for evaluation and use of automated scoring". In: *Educational measurement: issues and practice* 31.1 (2012), pp. 2–13

Demographic Fairness

In the years leading up to the Hewlett Foundation study, researchers at ETS suggest that bias is worth checking[368]. But in that work, they fail to actually provide a methodology for the evaluation, and they certainly never provide data. Years later, some of those same researchers did build and release an open source tool for evaluating fairness in automated scoring[369,370]. Yet again, though, they decline to actually provide any hard numbers on how their systems perform in their own audits, instead providing simulations. And while I single out ETS here, other vendors have also made no effort to publish fairness audits of their systems, and don't even mention that such a problem might exist. By that standard, in fact, ETS is doing the *best* among their peers by acknowledging such a problem may exist at all.

This is hard to find in the automated essay scoring community. In industry, product developers are averse to studying these problems directly. In many cases it is more prudent to not check for a problem too closely, to stay ignorant and uncertain of the effects of algorithmic decision-making on your students; for the alternative might be to dig in, find the inequity that you fear might be present, and then fail to gather the development resources and budget to fix the problem in a timely fashion. In that hypothetical snare, your company has moved from ignorance to *willful harm*, using algorithmic systems you know to exaggerate unfair outcomes for users, without a roadmap for correction. This is a truly unacceptable position to put a company

[368] David M Williamson, Xiaoming Xi, and F Jay Breyer. "A framework for evaluation and use of automated scoring". In: *Educational measurement: issues and practice* 31.1 (2012), pp. 2–13

[369] Nitin Madnani et al. "Building better open-source tools to support fairness in automated scoring". In: *Proceedings of the Workshop on Ethics in Natural Language Processing at ACL*. 2017, pp. 41–52

[370] Anastassia Loukina, Nitin Madnani, and Klaus Zechner. "The many dimensions of algorithmic fairness in educational applications". In: *Proceedings of BEA*. 2019, pp. 1–10

in, and so the questions remain largely unstudied. Over the years, I have tried to find an avenue for investigating questions of bias and equity in the context of automated essay scoring. But I have consistently struggled to get projects started in earnest; and when I did see things get underway, have had projects closed suddenly and prior to confirmation and publication of results. Outside of my own earlier work, which hints at these questions but never addresses it head-on, very few of the developers building automated scoring systems have been willing to ask questions about fairness in the tools they build.

In addition to essay text, I also have access to demographic information for students, including age, race, and gender. Among those essays, the population is 77% White and 58% women, with a median age of 32; a demographic breakdown by race and gender is presented in Table 17. In the 2017 DAACS data, the median student is mid-career, coming back to college in their 30s after a decade or more in the workforce. The courses that they are about to enroll in are fully online.

		All		Scored	
		#	%	#	%
Race	White	4803	77.0	412	78.3
	Black	725	11.6	53	10.1
	Hispanic	193	3.1	14	2.7
	Asian	184	2.9	17	3.2
	Native	96	1.5	11	2.1
	Multiple Races	240	3.8	19	3.6
Gender	Women	3672	57.5	295	55.1
	Men	2711	42.5	240	44.9

Table 17: Race and gender demographics for all essays and the subset of essays assessed by humans using rubric scoring.

Methods

For the fairness audit here, I discard most demographic data and subdivide the student population into four groups, intersecting two variables: race and gender. For race, I group together self-identified White students and compare to all other students, collectively referred to as persons of color, or "POC." In this data, this includes all people identifying as Black, Native American, Hawaiian Native, Alaska Native, Asian, and Hispanic, who collectively make up 23% of the data. For gender, I compare self-identified men to women. As is common in technological settings[371], limitations in the underlying data result in erasure of transgender identities, which therefore cannot be evaluated. Additional variables like age, military status, and first language status are also available for future secondary analyses.

[371] Os Keyes. "The misgendering machines: Trans/HCI implications of automatic gender recognition". In: *Proceedings of CSCW* (2018)

For all statistical significance tests, unless specified otherwise I perform χ^2 tests on frequency tables of the joint occurrence of score points and factor of interest. When calculating statistical significance per trait on the rubric, I apply Bonferroni correction, dividing p-values by 9 to account for multiple comparisons. For this and all following analyses of demographic fairness, when referring to automated scoring, I use the automated predictions of the best-performing of the three models above on each trait.

Results

As shown in Table 18, only small differences in population performance are observed in this dataset; those differences are not significant after correcting for multiple comparisons (prior to correction, significant differences are found in Organization-Transitions and Paragraph-Focus).

		White - Women	POC - Men	White - Men	POC - Women
Content	Summary	2.29	2.14	2.22	2.14
	Suggestions	2.28	2.12	2.15	2.23
Organization	Structure	2.52	2.51	2.62	2.42
	Transitions*	2.05	2.12	2.24	1.89
Paragraph	Focus*	2.42	2.45	2.57	2.23
	Cohesion	2.52	2.49	2.62	2.40
Sentence	Correctness	2.46	2.31	2.41	2.17
	Complexity	2.35	2.31	2.39	2.29
Conventions		2.35	2.33	2.31	2.32

Table 18: Breakdown of mean rubric scores, by race and gender intersection.

When I investigate whether automation is more or less accurate for different subgroups of students, there are no significant differences in accuracy for seven of nine traits. For two traits, however, there is a difference: In both `Organization-Transitions` and `Paragraph-Focus`, there is a statistically significant difference in accuracy for the best automated model; these results are presented in Figure 27. Specifically, I observe more accurate (and consistently lower) scores for people of color on the `Organization-Transitions` trait, and more accurate (and consistently higher) scores for White students on the `Paragraphs-Focus` trait. This difference sets up some of our explanatory questions in the chapters to come.

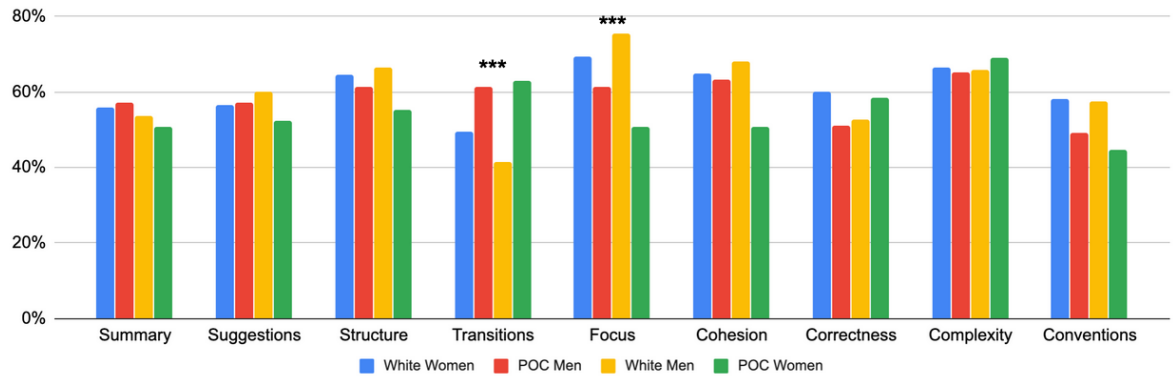


Figure 27: Accuracy of automated scoring by trait, broken out by race and gender.

New Data Collection

The basic results of this set of analyses were available to the DAACS team in summer 2019. They were about to embark on a new project: brick-and-mortar schools, filled with traditional-age college students.

In the 2019-2020 school year, for the first time the DAACS system moved from use in online-only universities with mid-career, nontraditional students, to a brick-and-mortar state school with a younger population of students, most of whom are coming directly out of high school. To improve the system for this use case, the team requires many changes in the system as they have developed it:

1. A change in training data, evaluating essays from traditional college-age students instead of mid-career adults.
2. A change in professional development as the system is used more proactively by academic advisors.
3. A change in expectations for retention, motivation, and on-time progress for their student population.

All of these changes have been happening in parallel, as part of a series of pilots and rollouts as the DAACS system matures. As part of this dissertation work I was able to assist a little. My work in auditing the original round of labeled data, and my experience in industry, revealed several areas for further refinement and improvement of the automated essay scoring system and the training data as it would be used in the new context. In particular, changes were made to the rubrics and the way that the development team interacted with the raters as they assembled the training set. A key portion of the DAACS work is describing whether changes in best practices can change the resulting model behavior. Our goal is not only to update the system to account for the shift in student populations, but also to intentionally correct for any warning signs that

are uncovered in the existing training set. The primary warning signs were the score distribution and the demographic shift.

First, on several traits in the original dataset, the full three-point scale was not used. On the Organization-Structure, Paragraphs-Cohesion, and Sentence-Complex traits, 5% or fewer of the training instances received a score of 1. In this newer round of data labeling, annotators were explicitly encouraged to use the full range of score points, avoiding any urge to "clump" in the middle of the score range for the sake of inter-rater reliability.

Next, content in the original corpus of essays came from primarily online students in their mid-30s. The topics they chose to write about was likely to differ from new texts from traditional brick-and-mortar institutions. Students in these life circumstances have different narratives and are likely to respond differently to the DAACS prompt. Domain transfer is not at all a solved problem in natural language processing[372], and so a new dataset is necessary in order to provide reliable scores for this new student population. Differences in age and context may also help our goal of explainable NLP, illuminating differences in student populations and what they choose to write about. Immediate research questions on this new data are:

[372] Sebastian Ruder et al. "Transfer learning in natural language processing". In: *Proceedings of NAACL: Tutorials*. 2019, pp. 15–18

- **RQ1:** Did human inter-rater reliability improve when conducting scoring in a more hands-on, local way?
- **RQ2:** Are the scores from the new human process equally predictable by machine learning methods?

Methods

We constructed a new set of 500 essays to use for building the updated dataset and training the updated model. This is composed of a subset of students directly from the University at Albany, who used DAACS as part of a pilot program on-campus, as well as a subset of students from the previous online domain. Those students who are explicitly in the "traditional" college age band of 18-22. Overall, because of this, the student population for this new dataset skews much younger than the original data. Race and gender representation was similar, as shown in Table 19.

The rubric was changed in several small ways:

- Additional detail was given on the `Content:Suggestions` trait. The new wording specifically clarifies the differences between tiers, to look for language of commitment in answers that receive a 3, while not expecting that commitment language at lower score points.

		Women		Men		All	
		#	%	#	%	#	%
Race	White	199	42.7	119	25.5	318	68.2
	Black	35	7.5	23	4.9	58	12.4
	Hispanic	25	5.4	23	4.9	48	10.3
	Asian	8	1.7	2	0.4	10	2.1
	Native	1	0.2	3	0.6	4	0.8
	Multiple Races	19	4.1	9	1.9	28	6.0

Table 19: Race and gender demographics for essays in the 2020 dataset. International students are not included in race statistics.

- In `Organization:Structure`, new language was added to make clearer when to give the lowest possible score.
- In `Paragraph:Focus`, additional language was added to encourage diversity in scores, especially around short essays with only one paragraph.
- In `Sentence:Correct`, explanation of the different score points was streamlined to make it clearer how to distinguish this category from conventions or usage errors for individual words, instead focusing on syntax errors like run-on sentences.
- The `Conventions` trait broken into two separate scores, as described in Table 20.

A major change was the shift in who actually conducted the rating and how they were trained. The initial round of data was labeled by a set of 11 raters who were trained remotely, then received guidance from follow-up sessions. This newer dataset, however, was much more hands-on. While most work was still performed remotely due to COVID-19 restrictions, raters had more regular check-ins and extensive guidance on labeling. Data was still double-scored, and in this iteration, raters were encouraged to discuss with the DAACS research team and with each other after most (but not all) batches of labeled data. This produced a much more time-consuming process, with much greater fidelity to the types of processes valued by rhetoric and composition scholars.

As a result of this much more conversational process for establishing inter-rater reliability, the DAACS team did not stay as distant from the rater discussion process as in the 2017 data. Instead, scores were assigned in small batches, maintaining independence between raters during scoring but with check-ins and discussion after each batch. The raters themselves noted in these meetings that this led to alterations in how to interpret rubric traits from batch to batch. The goal of this process was to improve overall quality of the rating at the cost of time and confidence from the raters. This differs from the process enacted at high-stakes scoring vendors, where raters go

Trait	Subtrait	Developing (1)	Emerging (2)	Mastering (3)
Conventions	Usage	Usage errors (such as incorrect word forms, subject-verb agreement, unaccountable shifts in POV) are numerous enough to distract a reader and/or interfere with meaning. Patterns of usage errors may be evident, suggesting that the writer lacks an understanding of basic usage rules and conventions.	Usage is generally correct. There may be errors but they are neither numerous enough nor serious enough to indicate that the writer lacks a basic understanding of the rules for usage.	Usage is correct. Usage errors, if any, are common and very minor.
Conventions	Punctuation	Errors in punctuation are numerous enough to distract a reader and/or interfere with meaning. Patterns of punctuation errors may be evident, suggesting that the writer lacks an understanding of key rules for punctuation.	Punctuation is generally correct. There may be errors but they are neither numerous enough nor serious enough to indicate that the writer lacks a basic understanding of the rules for punctuation.	Punctuation is correct. Punctuation errors, if any, are common and very minor.

Table 20: Changes to the Conventions traits in the 2020 revised rubric.

through a more standardized and hands-off calibration process that has remained largely stable for decades[373].

[373] Carol M Myford and Edward W Wolfe. "Detecting and measuring rater effects using many-facet Rasch measurement: Part I". in: *Journal of applied measurement* 4.4 (2003), pp. 386–422

Human Scoring Results

Human inter-annotator reliability differences between the two datasets are presented in Figure 28. In almost all traits, inter-rater reliability between humans has increased in the new domain with the new process, though the overall effect is small. The most important change between the two datasets was the change in distribution of scores. Whereas the 2017 data had multiple traits where scores of 1 were rare, appearing in 5% or fewer of all datapoints, across all traits no individual score point in the 2020 data occured in fewer than 7% of essays (exact distributions are given in Table 21). The DAACS team attributed this to a mix of lower-quality essays overall, making scores of 1 more frequent, as well as encouragement to raters to use the full scale, rather than focus on making "safe" scores that would artificially

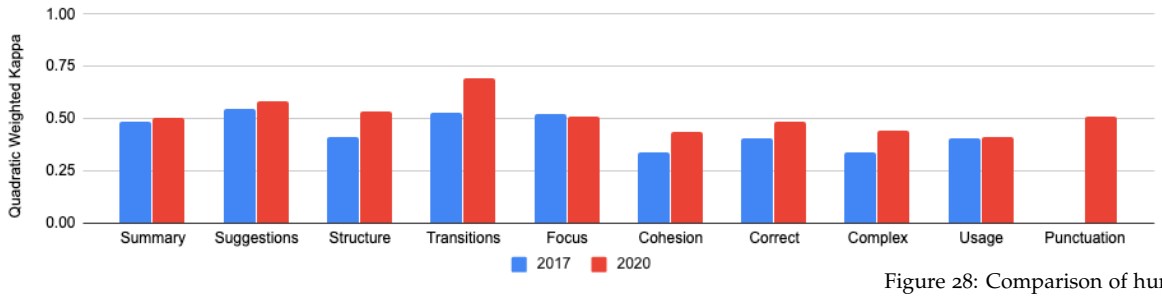


Figure 28: Comparison of human inter-rater reliability, in QWK, from 2017 to 2020 datasets, with changes made to rubric and process design.

		2017			2020		
Trait		1	2	3	1	2	3
Content	Summary	18.9	39.6	41.6	21.3	43.3	35.4
Content	Suggestions	22.6	33.3	44.2	17.5	32.3	50.2
Organization	Structure	4.6	35.5	59.9	12.6	43.3	44.1
Organization	Transition	20.9	46.8	32.3	39.2	32.7	28.0
Paragraphs	Focus	13.1	28.3	58.6	14.8	38.0	47.2
Paragraphs	Cohesion	3.5	38.6	57.9	7.3	37.0	55.7
Sentences	Correct	10.5	39.9	49.5	9.3	26.4	64.2
Sentences	Complex	5.0	54.7	40.3	9.1	49.8	41.1
Conventions	Usage	12.6	41.6	45.8	15.7	31.1	53.3
Conventions	Punctuation				21.1	44.9	33.9

Table 21: Percent distribution of each score point for each rubric trait, across datasets, in human scoring.

inflate inter-rater reliability.

Automation Results

The results of our machine learning pipeline are presented in Table 22. Here, the results are more mixed. While the model is able to effectively reproduce the high human inter-rater reliability on content traits, the effectiveness of the BERT-based neural model on sentence-level and conventions-level traits has all but vanished. Performance now drastically lags behind human inter-rater reliability. For comparison side-by-side with the 2017 results, Table 23 gives a detailed breakdown.

These results leave us with a mixed bag of conclusions about the shift in rater behaviors. The improvement in reliability for content traits is valuable, especially as the choice of content for these students will almost certainly be more impacted by domain transfer than changes in grammar or sentence-level traits. But the gap between automated and human scoring reliability on more low-level traits means that more work remains before these models can be used in a reliable way in the new domain. Future work might incorporate

data from both the 2017 and 2020 datasets into a single training set, especially on traits where the rubric criteria did not change materially. By nearly doubling the available data for a single trained model while broadening the range of essays available in the data, a better result might be achieved than the models trained on either 2017 or 2020 data independently.

	Human				Automation		
Trait	Exact	Exact+Adjacent	κ	QWK	Exact	QWK	QWK Gap
Summary	55.51	96.12	0.311	0.501	60.2	0.562	-0.061
Suggestions	64.90	95.31	0.426	0.58	59.18	0.442	0.138
Structure	60.41	97.76	0.359	0.535	58.16	0.374	0.161
Transitions	64.9	97.96	0.469	0.692	53.88	0.435	0.257
Focus	57.55	96.94	0.305	0.511	62.86	0.452	0.059
Cohesion	58.37	98.16	0.247	0.437	62.65	0.32	0.117
Correct	63.47	95.92	0.329	0.486	60.2	0.143	0.343
Complex	58.16	97.76	0.291	0.444	57.55	0.292	0.152
Usage	50.61	94.29	0.187	0.412	52.45	0.187	0.225
Punctuation	56.73	96.94	0.320	0.509	45.1	0.188	0.321

Table 22: Results of tuned models with new 2020 data.

	2017 Human	2020 Human	2017 Automated	2020 Automated
Summary	0.485	0.501	0.472	0.562
Suggestions	0.547	0.58	0.505	0.442
Structure	0.409	0.535	0.287	0.374
Transitions	0.526	0.692	0.247	0.435
Focus	0.522	0.511	0.559	0.452
Cohesion	0.338	0.437	0.319	0.32
Correct	0.403	0.486	0.377	0.143
Complex	0.338	0.444	0.366	0.292
Usage	0.403	0.412	0.336	0.187
Punctuation		0.509		0.188

Table 23: Comparison of automated reliability between 2017 and 2020 datasets.

Explaining Essay Structure

The previous chapter showed that levels of reliability approaching human-human agreement could be achieved through automated methods, using structural features, *n*-gram features, and neural methods. I used the optimized model from these experiments to build an automated essay scoring model that approached human inter-rater reliability on most traits.

I have argued the value of holistic explanation through quantitative but non-causal means; with Wikipedia deletion debates, I showed one approach for explaining model behavior with these goals in mind. But for DAACS, our explanation of model behavior in the standard NLP tradition has been narrowly on fairness, and group fairness, at that. So now, I will build on the preliminary findings from the group fairness audit and study a few specific research questions about student variation with an eye toward explaining where the model's decision-making comes from.

In the next two chapters I'll study the *style* of writing. Specifically I'll look at how the AES system makes predictions on essays containing recognizable behaviors from students, predictable writing "moves" that scholars in the field would recognize. I measure how these moves alter the scores students receive and the accuracy of the automated models; I also look at how an author's identity, as defined by their race and gender, shapes which of those choices they make.

Regardless of the literal instructions provided to raters or the calibration process for scoring, students may have textual characteristics in their writing that indicate the text was written by a person of a certain background, steeped in a certain set of cultural or academic traditions. These differences have the potential to influence rater scores, either human or algorithmic, and either subconsciously or intentionally. The presence or absence of such cues – especially those that indicate an academic, affluent background – may lead to differential outcomes for students even on a low-stakes assessment like the DAACS. This self-reinforcing tradition of measuring academic achievement is well-established in the critical theory literature[374,375]; but discussion of their impact on AES is virtually absent even after decades of research.

In this first explanatory chapter, we're going to study the structure of essays. Specifically, I'm zooming in on one genre form, the *five-paragraph essay* (hereafter, the 5PE), and the way it is scored by hu-

[374] Pierre Bourdieu. *Homo academicus*. Stanford University Press, 1988

[375] Sara Delamont, Odette Parry, and Paul Atkinson. "Critical mass and pedagogic continuity: studies in academic habitus". In: *British Journal of Sociology of Education* 18.4 (1997), pp. 533–549

mans and by machines. This structural form is widely taught in high school for the purposes of basic writing competency and test prep, beginning in the late 19th century[376]. It consists of an introduction paragraph with a thesis statement; three body paragraphs focusing on distinct topics, each providing exactly one support for the thesis; and a conclusion paragraph that summarizes the argument of the essay. This structure is consistently taught in United States high schools and subsequently un-taught in higher education[377]. By focusing on five-paragraph essays, this paper opens up an initial exploration of a new question about what standard AES models actually evaluate. Do they give reliable scores to both the texts that follow a formula taught by well-intentioned English teachers and test prep tutors, as well as more nontraditional texts that eschew that structure? My specific questions are:

[376] Matthew J Nunes. "The five-paragraph essay: Its evolution and roots in theme-writing". In: *Rhetoric Review* 32.3 (2013), pp. 295–313

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

- **RQ1:** How prevalent is the five-paragraph form in incoming first-year college student writing?
- **RQ2:** How are 5PEs scored when following traditional, rubric-based essay assessment?
- **RQ3:** Are 5PEs more or less prone to error due to automation, compared to other essay forms?

In the DAACS data, I'll construct a heuristic that identifies 5PEs automatically. With these essays identified, I measure pre-existing prevalence statistics. Then, I focus on a subset of essays in the dataset that were manually scored by trained educators on a rubric. I measure how these scores are associated with 5PE form.

An interaction between use of the five-paragraph essay form, race, and gender would not be surprising. Five-paragraph essays might show up in the essays of students from marginalized backgrounds. For schools struggling to maintain funding based on their results on standardized test scores, much of their focus on schooling comes in the form of intensive drill-based exercises, which "teaches to the test" by having students memorize specific structural elements of essay texts. This phenomenon disproportionately takes up instructional time for students attending low-income schools[378]. But the opposite may also be true: adherence to school expectations on assignments (and awareness of what those expectations even *are*) is associated with students from highly educated, affluent families[379], and girls in particular tend to outperform boys in writing assessments, an effect that increases with age[380]. This effect is driven at least partially by boys' demotivation toward academic achievement. Research suggests that success in educational tasks, especially those that are low-stakes

[378] Louie F Rodríguez. "Moving beyond test-prep pedagogy: Dialoguing with multicultural preservice teachers for a quality education". In: *Multicultural Perspectives* 15.3 (2013), pp. 133–140

[379] Ronny Högberg. "Cheating as subversive and strategic resistance: vocational students' resistance and conformity towards academic subjects in a Swedish upper secondary school". In: *Ethnography and Education* 6.3 (2011), pp. 341–355

[380] Caroline Scheiber et al. "Gender differences in achievement in a large, nationally representative sample of children and adolescents". In: *Psychology in the Schools* 52.4 (2015), pp. 335–348

and not tied to a specific prestige-based outcome, has socially constructed feminine connotations[381,382]. Investigating the question of how this plays out quantitatively in student writing is a gap in the AES literature, but even moreso, it is a gap in the *higher ed composition* literature more broadly. To my knowledge, no large-scale study has been conducted on the prevalence of five-paragraph essays among demographic subgroups in higher education.

I show for the first time the ways that use of this form differs at the intersection of race and gender, and that this shapes the demographic fairness results of the previous chapter. I'll then conclude the investigation by using the explanation to *mitigate* this bias, encoding awareness of the use of that formula directly in the machine learning model's features. The resulting representation, with more domain knowledge embedded, improves reliability of the models for all students. Furthermore, after the improvement, underlying racial and gender disparities in model reliability are no longer present.

[381] Stephen Frosh, Ann Phoenix, and Rob Pattman. "The trouble with boys". In: *The Psychologist* 16.2 (2003), pp. 84–87

[382] Andrew Wilkins. "Push and pull in the classroom: competition, gender and the neoliberal subject". In: *Gender and Education* 24.7 (2012), pp. 765–781

Methods

This dataset comes from live usage of DAACS. After completing these surveys, students complete a writing task, asking them to reflect on the content of their results and make plans for their upcoming college experience. The corpus consists of the training set from the previous chapter, as well as 5,712 essays submitted to this writing prompt in the DAACS platform. All essays were collected between April 2017 and February 2018. For all analyses of performance on this dataset, because human labels were not available, Essays were scored with the best-performing models from the previous chapter.

Finding Five-Paragraph Essays

Any results will rely heavily on knowing what a five-paragraph essay actually is, and automatically finding those essays in a large-scale corpus. To help readers grasp this intuitively, I provide examples of the actual essays described in this work in Table 24. These essays are divided into structural essays that use function words like "First," "second," and finally" to signal their five-paragraph structure, and topical essays that label each body paragraph with specific content topic words like "metacognition," "motivation," or "strategies."

To identify 5PEs automatically in the dataset, I define a generic function for pattern matching based on heuristic keywords. This function takes as input a paragraph p and two sets of keywords, the first set labeled `IN` and the second labeled `OUT`. Each keyword searches for stemmed, case-insensitive matches in p, which means

	Essay Text (first sentences from each paragraph)
Topical (exact)	¶1 My self-regulated learning survey was definitely an eye-opener for me. [...] ¶2 Metacognition is an area that I knew I had some strengths in but did not realize on how strong I am. [...] ¶3 Strategizing has always been one of my strong suits. [...] ¶4 Motivation is something that I have never lacked. [...] ¶5 This survey has enabled me to focus on my strengths to make sure that I use all the tools I have for myself to get through these courses proficiently. [...]
Topic (partial)	¶1 The human brain is a unique thing. [...] ¶2 Metacognition is defined, as the understanding of ones own thoughts. [...] ¶3 I am the first of my family to seek out an education higher than a high school diploma. [...] ¶4 Working on busy ambulances and staying awake for 24 hours every fourth day of my life has a way of motivating myself to improve my life that is indescribable. [...] ¶5 Using the focus of improving my time management and decreasing my anxiety toward education, I will be off the ambulance. [...]
Structural(exact)	¶1 The SRL assessment survey composed an evaluation of the strengths and weaknesses in self-regulated learning skills, additionally, it garnered recommendations on favorable strategies to develop self-regulated learning techniques. [...] ¶2 Metacognition is the first component of the SRL survey. [...] ¶3 The next important category in the SRL assessment survey was developing or enhancing particular strategies to foster a fruitful learning experience. [...] ¶4 The final section of the SRL assessment survey was motivation. [...] ¶5 In conclusion, the SRL assessment survey was a valuable insight to my learning abilities and rendered imperative strategies for success. [...]
Structural(partial)	¶1 The DAACS self-regulated learning survey results indicated that there are some areas in which I can improve, although I am, on the whole, a skilled learner. [...] ¶2 Planning is what I do, or should do, before I begin a study session or assignment. [...] ¶3 Monitoring is a second area of potential improvement. [...] ¶4 After the completion of an assignment or study session, it's valuable for me to engage in evaluation, another suggested realm of growth for me. [...] ¶5 The last recommendation I'd like to cover is managing time. [...] ¶6 In closing, I believe that the feedback provided by the DAACS self-regulated learning survey will help me to become both a more effective and a more efficient student. [...]

Table 24: Example first sentences of each paragraph from essays exactly or partially matching the 5PE heuristic search functions.

Table 25: Heuristic keywords used for matching five-paragraph essay components.

	IN	OUT
STRUCT1	first, begin	second, third, next, last, final
STRUCT2	second, next	first, begin, third, last, final
STRUCT3	third, last, final	first, begin, second, next
TOPIC1	metacognition	strategy, motivation
TOPIC2	strategy	metacognition, motivation
TOPIC3	motivation	metacognition, strategy

that any conjugation or word ending is a match; for instance, a search for [metacognition] would also match on [metacognitive] or [Metacognition]. I say a particular pattern matches p if both of the following conditions hold:

- p contains *at least* one instance of any term from the first set `IN`.
- p contains *exactly zero* of the terms from the second set `OUT`.

The exact set of keywords used for heuristic search are given in Table 25. For each essay in the dataset, I divide the text into paragraphs based on line breaks, and separate the text into the initial paragraph (the introduction), the final paragraph (the conclusion), and the middle "body" paragraphs. I use the pattern-matching function above to define two composite search heuristics over the body paragraphs: structural and topical.

- **Structural Matching:** For structural five-paragraph matching, I define an exact structural match as any essay where the STRUCT1, STRUCT2, and STRUCT3 matching functions are matched consecutively in the first three body paragraphs of the essay. I define a partial structural match as any essay where the three matching functions are matched, in order, in the body paragraphs, but with one or more additional paragraphs in between (allowing for essays that contain one or more interstitial or elaboration paragraphs for each topic).
- **Topical Matching:** I next look for topical five-paragraph essays, based on the three topics of the self-regulated learning survey that the students are asked to write about: metacognition, strategies, and motivation. I define an essay as an exact topical match if each of TOPIC1, TOPIC2, and TOPIC3 appears in exactly one body paragraph, and within that paragraph, the other two patterns do not appear. These topics match the three categories of results from the DAACS self-regulated learning results. I next say an essay is a partial topical match if the condition above is true for two of the three topic functions, and the third either never matches any paragraph in the text, or matches exactly two other paragraphs.

Demographics

I use the same demographic methods as in the first chapter on DAACS. I group together self-identified non-Hispanic White students and compare to all other students, collectively referred to as people of color, or "POC." In gender, I compare self-identified men to women. As in the prior analysis, this approach results in erasure of various identities including transgender students[383].

[383] Os Keyes. "The misgendering machines: Trans/HCI implications of automatic gender recognition". In: *Proceedings of CSCW* (2018)

Results

		Structural		
		No	Partial	Exact
Topical	No	74.6	0.3	0.5
	Partial	14.4	0.1	0.5
	Exact	8.7	0.1	0.8

Table 26: Overall prevalence of five-paragraph essays in the total dataset.

Prevalence and Impact of Five-Paragraph Essays

My first question asks how many students write five-paragraph essays unprompted when given an open-ended writing task in the online setting of DAACS. The results are displayed in Table 26. In total, 10.6% of essays are exact matches to one or both five-paragraph essay heuristics, and an additional 9.9% of essays are partial matches to one or both heuristics; collectively, five-paragraph essays make up one-fifth of all essays submitted to DAACS. This means that 79.5% of essays are not matched to either heuristic.

		No	Partial	Exact
Content	Summary***	2.14	2.44	2.61
	Suggestions	2.24	2.16	2.18
Organization	Structure***	2.47	2.70	2.91
	Transitions***	2.04	2.23	2.54
Paragraph	Focus***	2.36	2.71	2.82
	Cohesion	2.51	2.62	2.72
Sentence	Correctness	2.35	2.49	2.49
	Complexity	2.34	2.36	2.39
Conventions		2.34	2.32	2.32

Table 27: Mean score of essays in each category of five-paragraph form, marked with *** when there is a statistical significant relationship between form and score.

Table 27 shows the human scoring in the dataset, divided between essays that are exact, partial, and non-matches to the five-paragraph form; the same data is visualized in Figure 29. A significant relationship exists in four traits, though the magnitude of the relationship varies. Where the form is a significant influence on scoring,

we would expect to see a “stairstep" pattern: exact matches performing higher than partial matches, which then perform higher than non-matches. This exact pattern does appear across five traits (`Content-Summary`, both document-level organization traits, and both paragraph-level traits, though the pattern in `Cohesion` is not statistically significant after correcting for multiple comparisons). Adherence to the five-paragraph form has no significant relationship with `Content-Suggestions` scores, sentence-level traits, or grammatical conventions.

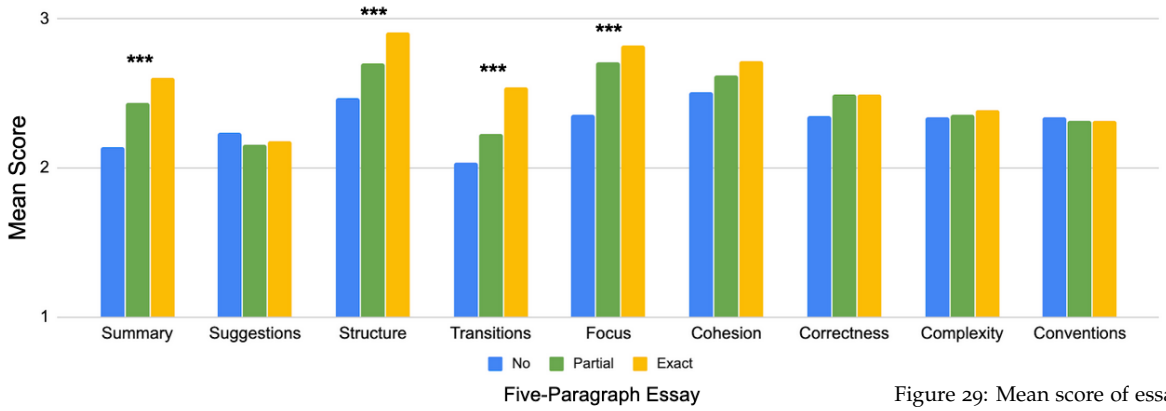


Figure 29: Mean score of essays in each category of five-paragraph form, marked with *** when there is a statistical significant relationship between form and score.

I next show the difference in exact accuracy of those automated scoring methods, divided between essays that are exact, partial, and non-matches to the five-paragraph form. My results match very closely to the results above: in Figure 30 a similar stairstep pattern appears in automation accuracy on the same traits. On `Content-Summary`, `Organization-Structure`, `Paragraph-Focus`, and `Paragraph-Cohesion`, automated accuracy is higher for 5PEs than for all other forms. In these cases, models are effective at recognizing the presence of indicators for five-paragraph essays, which are strongly associated with higher scores, and incorporating those signals into accurate reproduction of high scores. On `Organization-Transitions` and `Sentences-Correctness`, the opposite pattern is observed. Some other indicators of low scoring essays are more easily recognized, while essays following the five-paragraph form actually receive less accurate scoring from the machine learning classifier. No relationship between automation accuracy and the five-paragraph form is observed in the `Content-Suggestions`, `Sentences:Complex`, or `Conventions` traits.

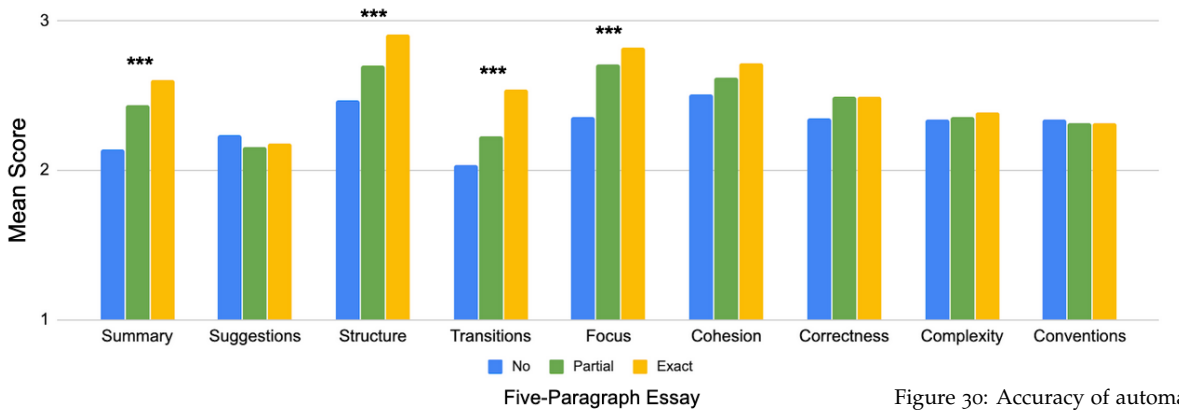


Figure 30: Accuracy of automated scoring by trait, broken out by 5PE form.

Essay Scoring Trends by Demographic Group

The following series of results attempts to measure whether the automation results above have any unequal distribution between individuals based on demographics of race and gender. I find that use of the five-paragraph essay structure is not distributed evenly across those demographics. As seen in Figure 31, there is no effect for men, but there is a highly significant effect for women, mediated by race ($\chi^2 = 26.7, p < 0.001$). White women are significantly more likely to follow the exact five-paragraph form compared to any other group, including a relative increase of about 30% compared to women of color.

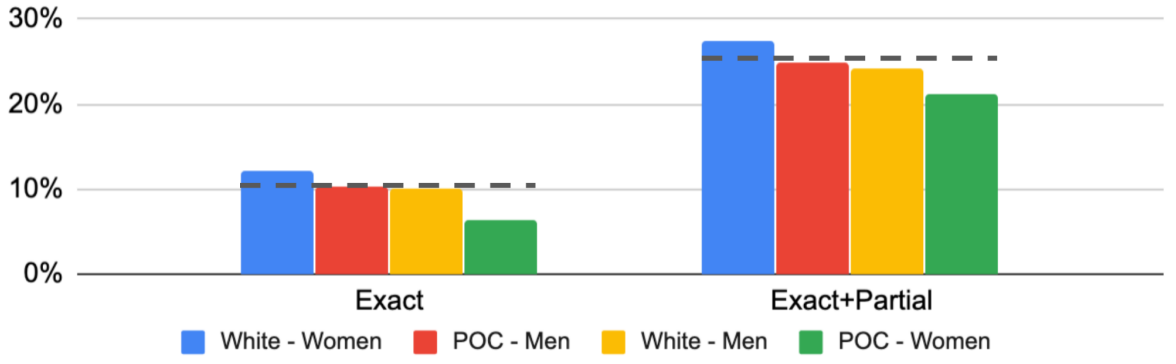


Figure 31: Breakdown of five-paragraph essay frequency by race and gender intersection. Dashed lines indicate whole-population frequency.

This difference in form does not correspond directly to a significant difference in overall performance by race and gender, as measured by rubric scoring. As shown in Table 10, only small differences in population performance are observed in the dataset; those differences are not significant after correcting for multiple comparisons

(prior to correction, significant differences are found in Organization-Transitions and Paragraph-Focus).

Summary of Results

To summarize our results so far:

- Use of the five-paragraph essay form is significantly associated with rubric scoring on document-level organization traits as well as paragraph-level cohesion and focus traits.
- While automated models generally are able to match human inter-rater reliability, there is a gap in reliability for AES when scoring document-level organization traits.
- The five-paragraph form appears disproportionately often in the essays written by White women, and disproportionately rarely in essays written by women of color; however, this difference does not correspond to higher scores for demographic populations.
- Racial discrepancy exists in the accuracy of automated scoring, on those same organization and paragraph traits. No such racial discrepancy for automated scoring is observed on content-level traits, sentence-level traits, or scoring for conventions.

Modeling the Five-Paragraph Essay

The data shows that the presence of 5PE structure does have an impact on automated writing assessment reliability. Next, I test how reliability changes when directly encoding features representing whether an essay follows the 5PE structure. In this final result for the chapter, I show that this explicit labeling of essays improves the accuracy of automated essay scoring. Furthermore, this improvement is shared across demographic groups, rather than being concentrated in any one subpopulation of students. The improvement is greatest in the assessment of document-level organization traits that were previously least reliable, and the resulting models no longer differ significantly in accuracy across demographic subgroups.

This leads us to question whether or not the accuracy of the automated essay scoring models can be improved specifically based on the patterns above, and the disparate performance of the models across race and gender mitigated. I attempt to do so by explicit modeling of the 5PE form.

In this experiment, for each automated scoring model, I make eight new quantitative features available. First, six count features are added, each measuring the number of paragraphs matching one

		Human	Automated (base)	Automated (5PE)	Difference
Content	Summary	0.473	0.472	0.465	-0.007
	Suggestions	0.532	0.505	0.508	0.003
Organization	Structure	0.392	0.287	0.353	0.066
	Transitions	0.516	0.333	0.403	0.07
Paragraph	Focus	0.519	0.559	0.601	0.042
	Cohesion	0.339	0.319	0.34	0.021
Sentence	Correctness	0.396	0.377	0.39	0.013
	Complexity	0.32	0.366	0.327	-0.039
Conventions		0.388	0.336	0.322	-0.014

Table 28: Reliability of automated essay scoring before and after 5PE encoding, in QWK.

of the six pattern-matching functions described in my initial five-paragraph essay extraction. Then, I add two categorical features indicating whether an essay is, overall, an exact, partial, or non-match for each of the structural and topical five-paragraph essay heuristic functions. My rationale for this inclusion is that explicitly informing the machine learning algorithm of whether an essay fits the five-paragraph form is beneficial; doing so allows the model to not only directly weight the appearance of those forms, but to stop giving weight to indirect proxy features that were stand-in evidence of those forms in the initial representation.

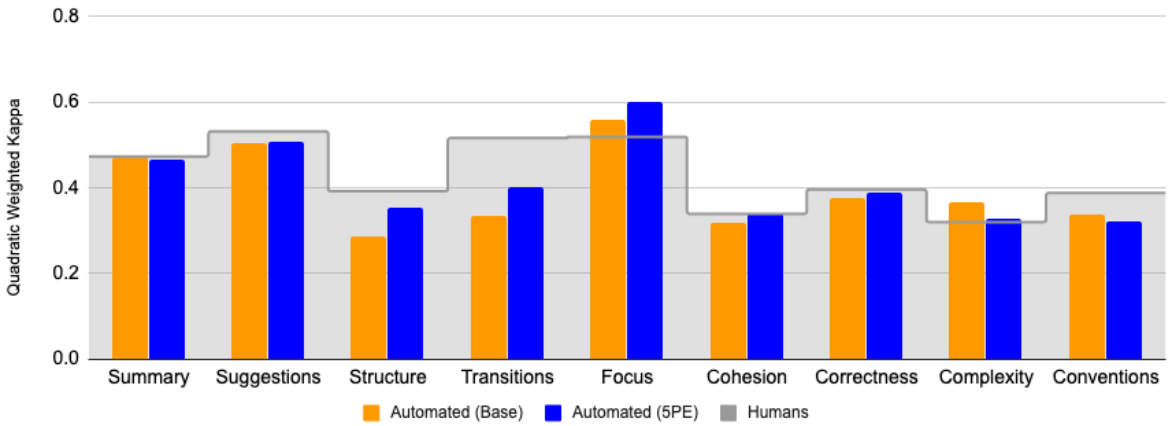


Figure 32: Reliability of automated essay scoring before and after 5PE encoding. Grey shaded area indicates human inter-rater reliability.

Table 28 and Figure 32 demonstrate the impact of encoding this information for the machine learning classifier. Performance improves for automation of scoring for three traits: `Organization-Structure`, `Organization-Transitions`, and `Paragraphs-Focus`. Not coincidentally, these overlap closely with the traits where essay scores were correlated with five-paragraph use in the original, human scoring data.

The demographic impact of this improvement in performance

			White Women	POC Men	White Men	POC Women
Organization	Structure	Base	64.6	61.2	66.5	55.4
		+5PE	66.8	55.1	69.1	63.1
	Transitions	Base	49.6	61.2	41.5	63.1
		+5PE	54.4	61.2	46.8	64.6
Paragraph	Focus	Base	69.5	61.2	75.5	50.8
		+5PE	71.2	65.3	72.3	56.9

Table 29: Accuracy of automated scoring, broken out by race and gender; only traits where reliability was improved by adding 5PE features are shown. In modified models, there is no longer any significant difference in accuracy by demographic subgroup.

is broken out, finally, in Table 29. Across the board, QWKs go up as the models are able to factor in the additional encoded information as part of their predictions; and after this modification is made, there is no longer any statistically significant difference in the accuracy of the automated models. For `Organization-Structure`, automated scoring is now within the threshold that Williamson et al.[384] recommend for reliability matching human judgment; only the `Organization-Transitions` dimension, with a gap of 0.12, is still unreliable using industry reliability norms for automated assessment.

[384] David M Williamson, Xiaoming Xi, and F Jay Breyer. "A framework for evaluation and use of automated scoring". In: *Educational measurement: issues and practice* 31.1 (2012), pp. 2–13

Explaining Essay Content

Obviously, an enormous amount of research has gone into automated scoring of student essays. But let's remember the context of most of that work: the essays are written to get high-stakes scores, measuring on rubrics for environments like the GRE, TOEFL, and in the United States, Common Core standardized testing. The current state of the research field is optimistic: when scoring student writing, composed in a timed environment, with a clearly defined and relatively objective rubric, automation is effective at reliably reproducing the decision-making that would occur at large-scale scoring vendors like ETS, ACT, Pearson, or the College Board.

Much of this research assumes student writing shares some basic, uniform characteristics. Students are assumed to be *school-savvy*[385], with a shared understanding of disciplinary norms around what kind of text they're supposed to write. They're also assumed to be *adherent*, complying with the instructions of the assignment and doing their best to receive a high score. To the extent that there is non-adherence, it's assumed that it's in the direction of over-performance due to cheating. The primary outcome variable, writing ability as measured by the test construct, is privileged above other dimensions of student variation. Researchers have relied on the belief that the noise of a student's technical literacy, motivation, or the circumstances of their home life will be drowned out by the signal of measurable writing skill, in the aggregate.

Perhaps this is an appropriate set of assumptions for high-stakes testing environments. But as AES and AWE systems move into environments like classroom instruction and academic advising, cracks begin to form. A large and growing body of research has shown that students are neither homogenous in content knowledge nor in goals and intentions when interacting day-to-day with education technology. Students and schools do not have aligned beliefs about technology use and, and the most significant aspect of school rules is not what they make you do but how they dictate norms of *how things should be done*[386]. Additionally, software developers and pedagogy scholars alike now acknowledge that students bring their whole self to school, belying attempts to measure students solely on academic skill[387]. Practically speaking, what this means is that NLP researchers have to recognize that non-adherent, even adversarial writing in education comes in many forms[388].

[385] Neil Selwyn. "Exploring the 'digital disconnect'between net-savvy students and their schools". In: *Learning, Media and Technology* 31.1 (2006), pp. 5–17

[386] Neil Selwyn and Scott Bulfin. "Exploring school regulation of students' technology use–rules that are made to be broken?" In: *Educational Review* 68.3 (2016), pp. 274–290

[387] Linda Darling-Hammond and Channa M Cook-Harvey. "Educating the whole child: Improving school climate to support student success". In: *Palo Alto, CA: Learning Policy Institute* (2018)

[388] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input". In: *Proceedings of NAACL*. 2018, pp. 263–271

In the financial sector, regulatory compliance prevents banking institutions from fully automating their relationship with customers, under a broad suite of "Know Your Customer" laws[389]. These laws require banks to be able to answer straightforward questions about their clients, like their identity and general purpose for their banking accounts, and also require ongoing monitoring of transactions in and out of accounts. Banks must know some bare minimum facts about the actual use of their automated systems. Education technology has no such requirement; in many cases, essays are written but never read, and whole corpora of student writing can be collected and trained without inspection or review for their contents, the identity of their authors, or the student's goals in interacting with our own fully automated systems. I believe this is a missed opportunity to recognize the "whole student" that interacts with automated systems in education[390].

[389] Philip J Ruce. "Anti-money laundering: The challenges of know your customer legislation for private bankers and the hidden benefits for relationship management (the bright side of knowing your customer)". In: *Banking LJ* 128 (2011), p. 548

[390] Nicholas Yoder. "Teaching the Whole Child: Instructional Practices That Support Social-Emotional Learning in Three Teacher Evaluation Frameworks. Research-to-Practice Brief". In: *Center on Great Teachers and Leaders* (2014)

Building on the last chapter's approach to five-paragraph essay analysis, this chapter goes wider, building more understanding of the actual text of the 2017 DAACS dataset. This investigation shows a wide variation in student response strategies and topic choices; a range of presuppositions of disciplinary norms and conventions for academic writing; and differences in who follows system instructions at all. Instead of a population of students varying primarily by traits that can be scored on an assessment rubric, the dataset instead consists of a wide spectrum of personal narratives and priorities. As a result, AES systems are subject to much more divergent inputs than is assumed by most machine learning and natural language processing research, calling into question the applicability of essentially all AES research in the transfer from standardized testing to other domains. In particular this investigation is structured to *explain automated scoring*, and so this produces the following research questions:

- **RQ1:** What topics/themes do students write about, and do they align to design expectations?
- **RQ2:** How do topics/themes affect the scores students receive from automated assessment?
- **RQ3:** How do topics/themes differ by student demographics, specifically race and gender?

Once these topics are explored for definition and labels that make sense, I turn to explanation based on the findings. I include related student behaviors motivated from the research literature on student behavior, like use of personal narrative and non-adherent behavior that indicates student disengagement. I show that these behaviors are closely tied to the topics that were automatically identified and

can tell us more, quantitatively, about the scores and distributions that we see in the dataset. My goal, as in the Wikipedia data, is to show that algorithmic decision-making can serve as a source of new insight, coming at the intersection of statistical evidence and more social science-driven observations and hypotheses.

Topic Modeling Methods

In the previous chapter, I found wide prevalence of the five-paragraph form; 20.5% of essays met a simple heuristic-based matching algorithm for identifying five-paragraph essays. These essays received significantly higher scores on four of the nine traits on therubric, concentrated in content, organization, and paragraph-level scores, while no relationship between the form and scores was observed in sentence-level traits or grammatical conventions. Furthermore, I found that these essays were concentrated in significantly larger numbers among essays by White women students. But up to this point, I haven't spent any real time describing *what the students wrote about*, beyond the structural constraints of the five-paragraph essay form. To fix this and make sense of the large set of student writing available to me, I turn to topic modeling.

Technical Approach

To perform topic modeling in a rigorous yet unsupervised way, avoiding manual annotation of each paragraph across thousands of student documents, I perform Latent Dirichlet Allocation topic modeling[391]. This approach is standard in the NLP community, and has been extended to many adjacent fields when performing large-scale text analysis[392,393]. LDA takes as input a set of texts and a predefined number of latent topics to discover, which is a hyperparameter that must be tuned for the specific purpose and domain. The model then infers a distribution of vocabulary terms for each topic, supposing that each input text should be composed of only a small number of topics. Topic modeling using LDA has no direct access to student variables like demographics or essay scores; student text is the only input. The model I describe in depth here has 20 topics; later in my analysis I describe the robustness of the explanation to changes in the number of topics.

In this analysis, I treat each *paragraph* as a separate input document, rather than each full essay. Using paragraphs as the base unit of analysis for student writing is consistent both practically for analysis, and qualitatively based on the structure of student writing. By calculating topic distributions on a per-paragraph rather than per-

[391] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022

[392] Ashraf Abdul et al. "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda". In: *Proceedings of CHI*. 2018, pp. 1–18

[393] Qian Yang, Nikola Banovic, and John Zimmerman. "Mapping machine learning advances from hci research to reveal starting places for design innovation". In: *Proceedings of CHI*. 2018, pp. 1–11

essay level, I am able to assign more specific labels to subsections of student text in theanalysis. At even smaller granularities like sentences, however, vocabulary sparsity makes topic labeling imprecise and difficult to evaluate or interpret.

Naming Topics

Topic modeling is an inherently mixed-methods process, as each output topic is a probability distribution over vocabularies, and topics do not come with labels but must be interpreted based on subject matter expertise.

To determine the appropriate label for topics, I calculate standard saliency metrics from Chuang et al.[394] to find key terms that appear primarily in only one topic, and additionally performed qualitative review of the text of paragraphs that were labeled with a given topic.

Future work may improve on this method by working with expert raters to assign names to topics; we did this, for instance, in my previous work with Diyi Yang on cancer support communities[395]. In Part IV of this dissertation, I discuss what this approach might look like as a means of further strengthening the explanatory strength of an analysis. For the purposes of this chapter, though, no external annotators were involved.

[394] Jason Chuang, Christopher D Manning, and Jeffrey Heer. "Termite: Visualization techniques for assessing textual topic models". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 2012, pp. 74–77

[395] Diyi Yang et al. "Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities". In: *Proceedings of CHI*. ACM. 2019, p. 344

Topic Results

Structural Paragraphs

Three topics emerged that were typically used as structural paragraphs in five-paragraph essays. Of these, one represented paradigmatic `Structure:Introduction` paragraphs, and two represented paradigmatic "conclusions".

The two topics that were extracted as conclusions show very different stylistic approaches to closing off an essay. The first, which I label `Structure:Conclusion:Excitement`, emphasizes the following terms: *excited*, *new*, *journey*, and *forward*. Concluding statements like the following were typical of this style:

> *Being able to clearly see the way I learn, and the ways I can improve my learning is a great way to get off on the right foot. I feel more confident then ever to get started on my WGU journey, thanks to the results and feedback I received from the SRL.*

The alternative topic I label `Structure:Conclusion:Commitment`. These essays focused not on enthusiasm but on specific goal-based reasons to complete their degree; keywords included *degree*, *bachelor*, *earning*, *goal*, and and *dream*. while this is a motivating factor for

college enrollment in general, diploma completion was not an explicit topic of the essay writing prompt itself, making the salience of this theme in conclusion paragraphs noteworthy.

> *I am really looking forward to completing my classes and gaining a degree. This would be a major accomplishment in both my personal life and professional career. This could take me a lot further in life as well as be supporting my family.*

Body Paragraphs

The DAACS SRL survey results are the primary topic of essays written in thisdataset; these results are structured into three broad topic areas, each of which is broken down into 3-4 subsections for students to browse. The previous chapter's results showed a high prevalence of five-paragraph essays that used this structure from the user interface as a proxy for how they should structure their work, with one body paragraph for each section. One question that I anticipated going into this analysis was whether those sections and subsections would be identified as latent topics by LDA. The menu, as displayed to students, is shown in Figure 33.

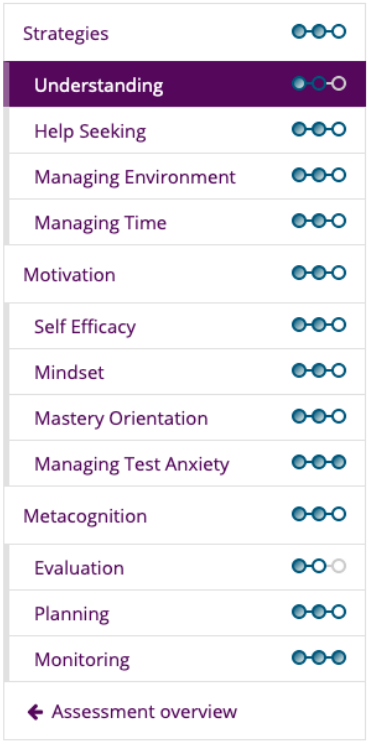


Figure 33: Sidebar menu for the DAACS self-regulated learning survey, which organizes results into a hierarchy.

I rapidly identified three topics that each corresponded to a general overview of one of the three main sections of the SRL survey – the kind of one-paragraph summaries typical of five-paragraph essays in the previous chapter. I labeled these topics as `Strategies:5PE`, `Motivation:5PE`, and `Metacognition:5PE`. Representative paragraphs labeled with these topics, for instance, look like this:

> *The last category was Motivation which I also scored in the high-range for. The category was broken down into four sections which were; self-efficacy, mastery orientation, anxiety, and mindset. My results indicated that I am confident in my learning abilities, that I find learning enjoyable, keep my anxiety levels low, and exhibit a growth mindset. These skills all contribute to my motivational level and overall capabilities to be a successful student.*

> *The next skill set is strategies which I have a good handle on. Managing environment and seeking help were were two categories that I know that I have a good grasp on. I was worried about time management and understanding. The survey shows that I have good skills in those areas. Even though I have strong skills in those area It still shows you what you can do to improve or stay strong in those areas.*

Beyond the three generic topics, eight additional topics matched closely to subtopics within the SRL survey structure. Three of the four subcategories for strategies – `Strategies:Help-Seeking`, `Strategies:Time`, and `Strategies:Environment` – were easily identifiable. Topics identified three of the four subsections in Motivation – `Motivation:Goals`, `Motivation:Mindset`, and `Motivation:Anxiety` – but no separate

topic emerged for self-efficacy. Two topics identified sub-areas for metacognition, one specific to `Metacognition:Planning` and one that grouped together the other two topics, `Metacognition:Monitoring`. An example paragraph, labeled with the help-seeking topic, looks as follows:

> *Finally, asking for help. After speaking with my mentor today, he strongly suggested I utilize "course mentors" for any and all subjects. These mentors can help save time with my studies, and provide very good feedback. Having subject experts at my disposal is a very powerful strategy, and one which I plan on taking advantage of.*

Between overview topics and subsection-specific topics, these identifiable, section-based topics make up 11 of the 20 topics that were surfaced by LDA; after accounting for the three structural topics, this leaves six topics that do not adhere closely to a structural part of the five-paragraph essay or the subject matter of the prompt.

Non-Adherent Topics

The three smallest topics by volume all related to various types of personal backstory and narrative for students. Less cohesive than the other topics related to specific five-paragraph essay themes, I group these three topics together as a single unit of analysis, `Narrative:Past`. One additional topic, labeled `Narrative:Future`, was also narrative-driven, but rather than focusing backward, looked forward to specific actionable plans for enacting strategies from the DAACS suggestions. Here, students described their intent for their upcoming college experience. While they do not necessarily adhere to a topic-focused essay form, itself they are clearly good-faith attempts to respond to the prompt. An example is:

> *I am currently working on my initial orientation for Western Governors University. I'm proud of the fact I have the opportunity to further my education as well as my role at work. The last time I attended college was about ten years ago. I attended a local private junior college, I felt I didn't have the confidence I needed. I attended everyday and completed all the course work that was assigned to me. I gradated with higher grades than I ever dreamed of. My goal is to become a dedicated student for the next two years while attending WGU.*

The model finally discovered two behavioral categories that did *not* correspond to topics from the survey, structural paragraphs from the five-paragraph form, or personal narratives. These represent a break from the "school-ratified" norm of the previous two subsections. Instead they represent transgressive or non-adherent behavior. The first category, which I refer to as `Non-adherent:Pasting`, identified paragraphs that were copied from elsewhere in the DAACS interface but were not the writing prompt itself. Pasting in prompt

text as part of a student essay is not expected behavior in any AES system's training data; its relatively high prevalence in real-world writing is noteworthy as a structuring strategy for student writers.

The final topic is perhaps the most confrontational and least adherent to "school rules." In this topic, the system identified a pattern of students directly confronting the system itself, describing the DAACS assessment in metalanguage, either for privacy, academic preparedness, or a variety of other reasons. One typical example of this `Non-adherent:Skeptic` topic is given below:

> *I find that no matter what the results of the test report, they cannot factor in all the required information that is to say that the questions are lacking in substance. The test has flaws in that the person taking the test is answering questions based on feeling; there is no real guarantee that questions answered, are honest. If the person taking the test is factoring in other information that makes the exception to a rule then answers will skew the results.*

To conduct further analyses by paragraph, for each topic, I assigned a value to each paragraph between 0 and 100, representing the integer percentage of tokens that were drawn from that topic by the LDA model. This produces a distribution over topics for each paragraph. I assigned category labels to paragraphs based on the topic that the plurality highest score, where the three smallest topics were combined into a single label. A full breakdown at the paragraph level for these eighteen topics of analysis, after grouping the three smallest topics, is visualized in Figure 34.

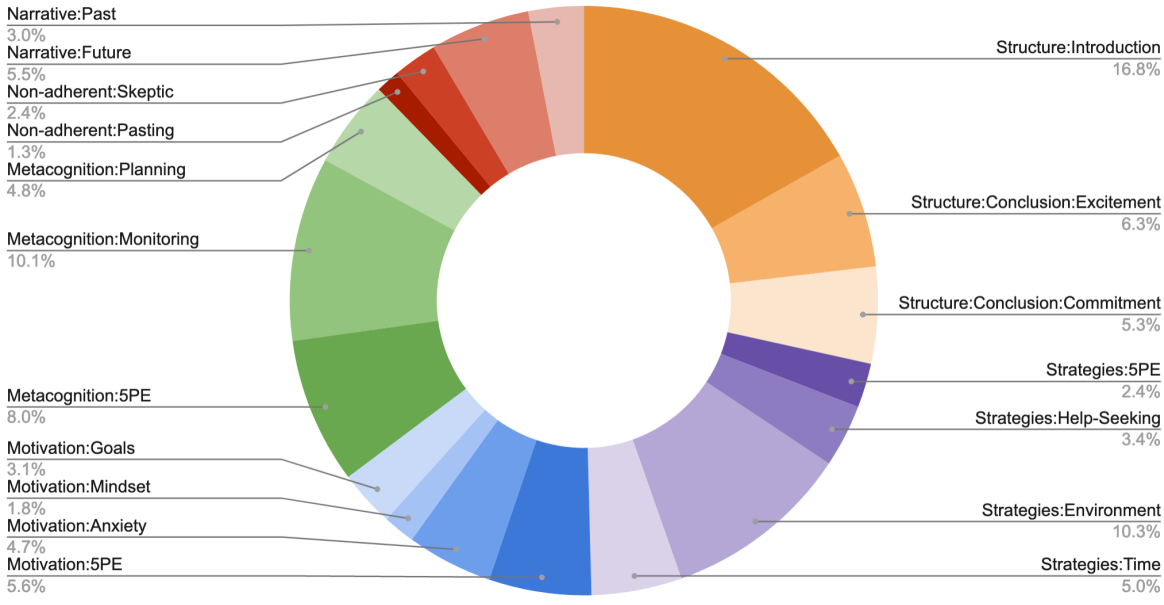


Figure 34: Distribution of topic assignments to paragraphs from the LDA model.

Topic	Subtopic	1	2	3	4	5	6	7	8	9
Structure	Introduction	+0.16	+0.16	+0.17	+0.20	+0.29	+0.16	+0.10	+0.08	+0.07
Structure	Excitement		-0.18							
Structure	Commitment	-0.21	-0.23							
Strategies	5PE									
Strategies	Help-Seeking									
Strategies	Environment		+0.16	+0.09	+0.09	+0.17	+0.07			
Strategies	Time									
Motivation	5PE	+0.46	+0.21	+0.22	+0.21	+0.29	+0.16			
Motivation	Anxiety									
Motivation	Mindset									
Motivation	Goals									
Metacognition	5PE	+0.38	+0.25	+0.21	+0.25	+0.30	+0.21			
Metacognition	Monitoring		+0.12			+0.16				
Metacognition	Planning		-0.27							
Narrative	Past									
Narrative	Future		+0.24	+0.20	+0.24	+0.30	+0.16			
Non-adherent	Pasting									
Non-adherent	Skeptic		-0.58							

Table 30: Differences for intersections of topic and trait, from mean population scores. Only significant effects are shown.

Topics and Scores

Next I study the intersection of topic appearance in essays with scores that those essays receive from automated essay scoring. These scores are the output of the trained model as described at the end of the previous chapter, including five-paragraph essay features. For this analysis, I am *only* interested in automated scoring, not human judgment, so I include only the data that was automatically scored, not data points or labels from the training set. I include one row for each of the eighteen topics for analysis. Because of the very large number of possible comparisons (18 topics rows and 9 trait sub-scores), I apply Bonferroni correction for multiple comparisons.

Results

The significant relationships between topics and automated trait scores are presented in Table 30.

The single most *consistent* significant relationship is the presence of an identifiable introduction paragraph. Essays where at lest one paragraph is labeled with the `Structure:Introduction` topic receive significantly higher scores on all nine traits – the only topic for which this is true. A few of the topics focused on specific body paragraphs predict stronger scores on high-level scoring traits. In particular, es-

says that include paragraphs labeled as part of five-paragraph essays in both motivation and metacognition receive higher scores. In strategies, those overview paragraphs are not the significant predictor of higher scores; instead, essays containing paragraphs labeled as focused on environment management are the subtopic associated with significantly higher scores. Outside of body paragraph topics, paragraphs labeled as `Narrative:Future`, typically reflecting commitment to future plans, receive significantly higher scores on five out of nine traits.

Meanwhile, the single *largest* effect is for essays containing paragraphs that are labeled as `Non-adherent:Skeptic`. These essays receive lower scores specifically on the trait measuring how well students expressed plans for following suggestions from DAACS.

Topics and Demographics

We can now move on to study how these behaviors differed by the demographic groups of the students in the dataset. I will divide these findings by topic subsets, using the same groupings as before (structural, body, narrative, and non-adherent topics).

	White		Black		Hispanic		Asian		Native		Multiple	
	W	M	W	M	W	M	W	M	W	M	W	M
#	2535	1854	428	243	88	91	81	86	53	32	130	91
%	44.4	32.5	7.5	4.3	1.5	1.6	1.4	1.5	0.9	0.6	2.3	1.6

Table 31: Race and gender counts for essays in the DAACS dataset, specifically among *unscored* essays.

I use the same demographic groups as in previous chapters on DAACS; on the larger unsupervised dataset, distribution of demographic labels are given in Table 31. I group together self-identified non-Hispanic White students and compare to all other students, collectively referred to as people of color, or "POC." In gender, I compare self-identified men to women. As in the prior analysis, this approach results in erasure of various identities including transgender students[396]. For each of the eighteen topics, I perform a χ^2 test of observed use of the topic in essays for the binary split by race and by gender, then for each of the four intersecting subpopulations. Bonferroni correction was applied to correct for multiple comparisons.

[396] Os Keyes. "The misgendering machines: Trans/HCI implications of automatic gender recognition". In: *Proceedings of CSCW* (2018)

Each results section will begin with a visualization of the relative difference in occurrences of documents where each topic appears, compared to the overall population mean. Statistically significant differences ($p < 0.05$ after correction), will be indicated with asterisks.

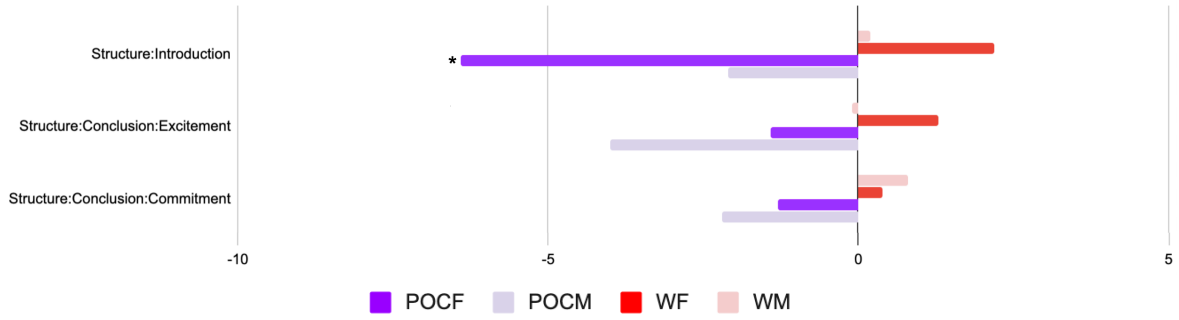


Figure 35: Subgroup differences for document structure topics.

Race and Structure

Results for the three topics related to five-paragraph essay structure are visualized in Figure 35. The largest gap is present for introduction paragraphs, which students of color were significantly less likely to use – and again, significantly less likely for women of color specifically compared to men. These results align with my prior findings that White women were most likely to follow standard five-paragraph essay forms. Small racial effects were seen for both types of conclusion topic, but the differences were not significant.

These results align with the hypothesis of a specifically racialized gap in use of a standard introduction paragraph, which is then, from the previous section, associated with lower scores. The educational system favors pupils who have academically educated parents, students that are socially and culturally close to school culture[397]. In writing, genre norms are a dynamic and locally mediated idea, not an "unmoving, absolutely knowable rule"[398]; here I find a place where scores are defensible on a rubric and fairly applied by those guidelines, yet lead to disparate racial outcomes based *not* on bias but on normative definition of what makes a good essay.

[397] Guangwei Hu and Jun Lei. "Chinese university students' perceptions of plagiarism". In: *Ethics & Behavior* 25.3 (2015), pp. 233–255

[398] Margaret Price. "Beyond" gotcha!": Situating plagiarism in policy and pedagogy". In: *College Composition and Communication* (2002), pp. 88–115

Gender, Race, and Topics

Results for body paragraph topics related to the SRL survey subsections are presented in a large graph in Figure 36. Several topics, including all three five-paragraph essay body paragraph topics, show no significant effects, and are not discussed in this section.

One of the most striking and highly significant variations focused on strategies for time and environment management. Men, and white men in particular, were significantly more likely to include time management paragraphs in their essays. By contrast, White women in particular were significantly more likely to discuss environment management. When I investigate this difference, the contents of the essay reflect a much more frequent focus on family life, childcare,

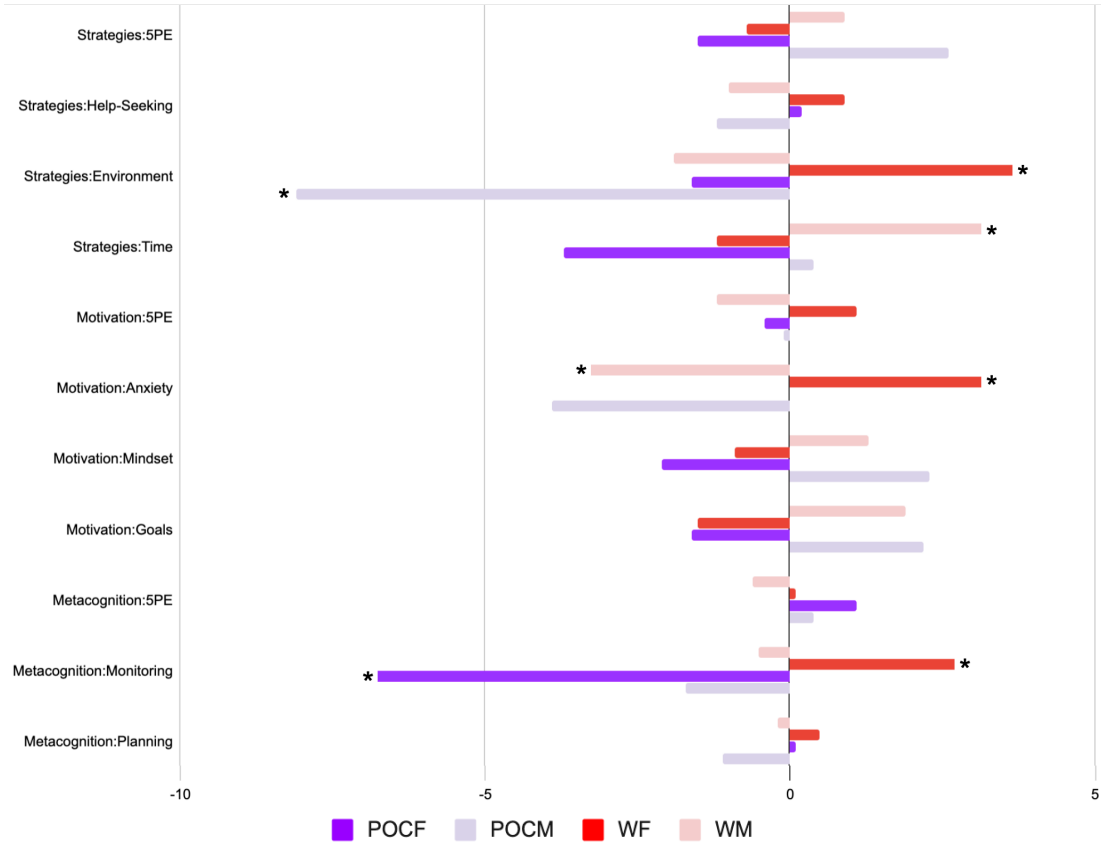


Figure 36: Subgroup differences for body paragraph topics.

and coordination with husbands and work responsibilities:

> *Another subject the survey brought up that I struggled in was managing my environment. As I talked about before, my husband and I already discussed having uninterrupted study time every day. During this time, my husband will be responsible for the kids for me to utilize that time I have set aside every day. This will be a learning process for us all, especially my daughter who is 7, but we will manage. I especially respect the suggestion of turning off or silencing my cell phone and other technology. It is too easy to get distracted and distraction is just wasted time.*

There was no such pattern of men talking about childcare or coordination with their wives to make space for degree completion.

In the Motivation category, a significant difference by gender was observed for discussion of anxiety. White women were significantly more likely than any other subgroup to talk about their anxiety in taking tests or returning to school, and men are significantly less likely. When students wrote about the topic of anxiety, they did so in good faith based on survey feedback:

> *There is nothing that I want more than to pass all my classes and keep moving to graduate. I will always stay motivated to do my best by staying positive and using the relaxing techniques like the survey suggested.*

If I plan things out more, or do my work at a steady pace, I may have less anxiety for a test. It also talks about practicing relaxation techniques. They talk about breathing exercises that may help to calm before a test. I have done these breathing exercises before and I do feel they are very relaxing. I also listen to some music before a test to reduce my anxiety and so I have a clearer head.

Finally among topics related to body paragraphs, writing about Monitoring is significantly less common among women of color, but significantly more common among White women.

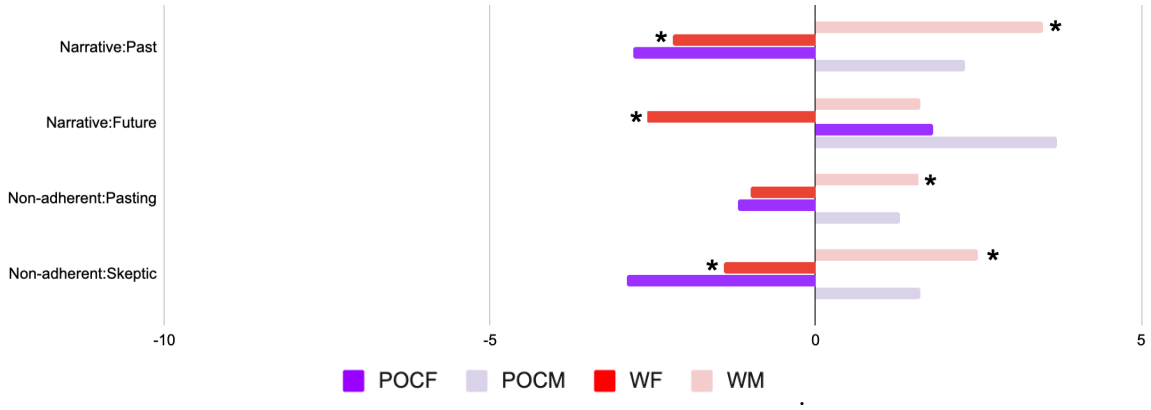


Figure 37: Subgroup differences for non-adherent paragraph topics.

Gender and Non-Adherence

The final set of results are displayed in Figure 37. Narrative writing that does not follow a topic-based five-paragraph form is significantly less common among White women, both for the past-oriented and future-oriented narratives. Backstory paragraphs were significantly more common among men, and White men in particular. Finally, non-adherent topics were significantly more likely to occur in essays written by White men, both for non-adherent pasting of prompt text, and especially for metalanguage expressing skepticism of the task itself.

This finding is complex, and cannot be boiled down to labeling White men as cheaters; Hogberg[399] argues that writing behavior is more about students' intertextual practices than their morality (especially based on differences in moral frame of reference and conceptions). Transgressive writing behaviors are again social, defined according to social expectations; a single judgment of whether non-adherent behavior is appropriate or inappropriate masks the agency that students feel in their use of educational interventions[400]. Given this, we must look to additional work in sex-role socialization theory to describe and explain this finding. That work argues that women are socialized to obey conventional norms[401]. In education, male stu-

[399] Ronny Högberg. "Cheating as subversive and strategic resistance: vocational students' resistance and conformity towards academic subjects in a Swedish upper secondary school". In: *Ethnography and Education* 6.3 (2011), pp. 341–355

[400] Erik Borg. "Local plagiarisms". In: *Assessment & Evaluation in Higher Education* 34.4 (2009), pp. 415–426

[401] Carol Gilligan and Ina Different Voice. "Psychological theory and womens development". In: *Cambridge, MA* (1993)

dents plagiarize more than female students[402]; more broadly, men are more okay with small, seemingly inconsequential academic offenses[403]. In short, *"male role norms are characterized by greater tolerance of minor transgressions."* Mac an Ghaill was an early proponent of the now more-developed evidence that boys reject schoolwork as feminine busywork[404], demotivating academic adherence as in conflict with performative masculinity in educational settings[405]. Being willing to "go along with" educational technology systems correlates with student learning outcomes[406], which ties to the broader findings from DAACS earlier in my work. This effect is also gendered, varying not only by student gender but by the perceived gender of the automated system as well; female-presenting agents are more abused than male agents[407], with threats and even violent or sexual content[408].

My results are also backed up by a broader set of findings outside of schools. Non-adherence is also predicted based on results outside of education: when users interact with an algorithmic system. Drivers for Uber and Lyft, for instance, believe that the system is efficient, but low levels of transparency drives users to work together to learn to resist and abuse it[409]. Drivers engage in a variety of behaviors to resist and game the system[410]. It would not be surprising to see similar behavior emerge in the educational setting.

Further Explanations from Essay Topics

The descriptions above corresponding to each topic were based on qualitative methods: a thematic analysis of reading the text of paragraphs as well as an inspection of salient words. But several of the claims that I make about the topics are testable quantitatively as well. So this section proceeds with a followup analysis on several specific claims that I make about the topics.

Position in Text

In my description of structural paragraph topics, I claimed that the primary purposes of paragraphs with these topics was to start or end five-paragraph essays.

By labeling the distance from the beginning and end of each essay, I confirm this interpretation in Table 32. For paragraphs I labeled as introductions based on the output of topic modeling, 64.1% appeared as the first paragraph in the document they were drawn from. Only one other topic, non-adherent pasting, appeared as the first paragraph more than 30% of the time; and in the cases where pasting occurred in the essay-initial position, students often used a restatement of the prompt as a structuring tool for writing their essay. For

[402] Bernard E Whitley, Amanda Bichlmeier Nelson, and Curtis J Jones. "Gender differences in cheating attitudes and classroom cheating behavior: A meta-analysis". In: *Sex Roles* 41.9-10 (1999), pp. 657–680

[403] Jean Underwood and Attila Szabo. "Academic offences and e-learning: individual propensities in cheating". In: *British Journal of Educational Technology* 34.4 (2003), pp. 467–477

[404] Máirtın Mac an Ghaill. "'What about the boys?': schooling, class and crisis masculinity". In: *The Sociological Review* 44.3 (1996), pp. 381–397

[405] Ursula Kessels et al. "How gender differences in academic engagement relate to students' gender identity". In: *Educational Research* 56.2 (2014), pp. 220–229

[406] Amy Ogan et al. "Oh dear stacy!: social interaction, elaboration, and learning with teachable agents". In: *Proceedings of CHI*. ACM. 2012, pp. 39–48

[407] Annika Silvervarg et al. "The effect of visual gender on abuse in conversation with ECAs". In: *International conference on intelligent virtual agents*. Springer. 2012, pp. 153–160

[408] DA Angeli, Sheryl Brahnam, and Peter Wallis. "Abuse: The darker side of human computer interaction". In: *Interact 2005*. 2005, pp. 91–92

[409] Min Kyung Lee et al. "Working with machines: The impact of algorithmic and data-driven management on human workers". In: *Proceedings of CHI*. 2015, pp. 1603–1612

[410] M Möhlmann and L Zalmanson. "Hands on the wheel: Navigating algorithmic management and Uber drivers'". In: *Proceedings of the International Conference on Information Systems*. 2017

Topic	Subtopic	Median Distance From Beginning	Median Distance From End
Structure	Introduction	**0**	4
Structure	Excitement	3	**0**
Structure	Commitment	3	**0**
Strategies	5PE	2	3
Strategies	Help-Seeking	3	2
Strategies	Environment	2	2
Strategies	Time	2	2
Motivation	5PE	3	1
Motivation	Anxiety	3	1
Motivation	Mindset	3	1
Motivation	Goals	3	1
Metacognition	5PE	1	3
Metacognition	Monitoring	2	2
Metacognition	Planning	1	3
Narrative	Past	3	3
Narrative	Future	3	2
Non-adherent	Pasting	1	4
Non-adherent	Skeptic	2	2

Table 32: Location of labeled paragraphs within essays, by distance from the beginning and end of the text. Highlighting in blue represents topics where the median appearance is at the beginning or end of essay texts.

the two topics I labeled as conclusions, 65.6% and 57.8% appeared as the final paragraph, while no other topic appeared in the final paragraph more than 30% of the time. The median distance from the beginning was 0 for introduction-labeled paragraphs, and the median distance from the end was 0 for conclusion-labeled paragraphs; this was not the case for any of the other topics measured.

Body Paragraphs in 5PEs

I next test whether the paragraphs that I label "five-paragraph" based on their contents actually appear disproportionately in five-paragraph essays, relative to the other topics. As seen in Table 33 and Figure 38, this is the case. The topic for introduction paragraphs and the three summary or overview topics are each the four topics most likely to appear in five-paragraph essays as labeled in the previous chapter, while the `Skeptic` topic representing non-adherent writing is least likely to appear in essays matching the five-paragraph form.

Recognizable Non-adherent Behaviors

One key outcome of this topic-based investigation was insight into the variety of *non-adherent* responses, texts that did not follow the standard essay genre that is a mainstay of student writing, especially at the high school and early college level. Students did not always

Topic	Subtopic	Exact	Exact+Partial
Structure	Introduction	29.09	56.22
Structure	Excitement	21.41	44.03
Structure	Commitment	19.06	44.49
Strategies	5PE	31.69	57.75
Strategies	Help-Seeking	19.95	45.62
Strategies	Environment	25.90	51.42
Strategies	Time	16.71	44.81
Motivation	5PE	42.44	71.37
Motivation	Anxiety	19.05	46.35
Motivation	Mindset	22.33	55.09
Motivation	Goals	24.13	52.28
Metacognition	5PE	38.34	68.24
Metacognition	Monitoring	21.25	46.88
Metacognition	Planning	21.09	46.49
Narrative	Past	19.10	41.89
Narrative	Future	25.45	53.12
Non-adherent	Pasting	28.57	48.38
Non-adherent	Skeptic	16.27	39.37

Table 33: Percentage of paragraphs labeled with each topic that appear in exact- and partial-match five-paragraph essays.

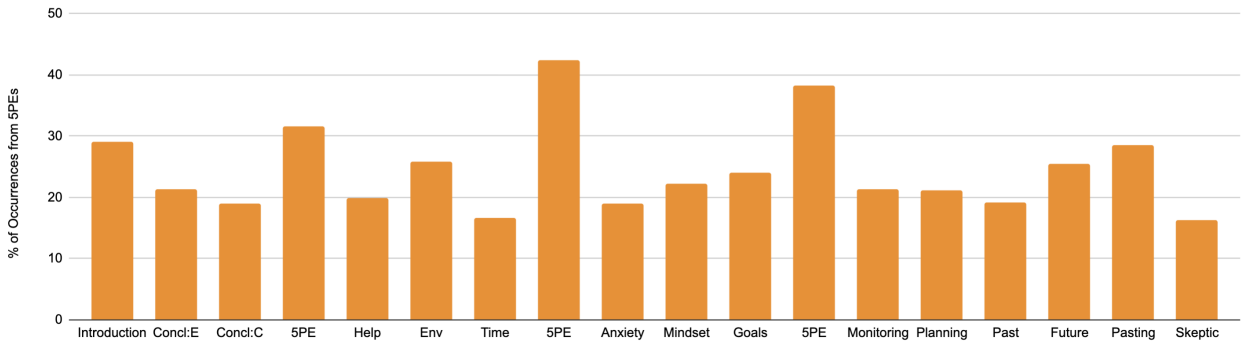


Figure 38: Data from Table 33, including exact matches only.

hew closely to the instructions or expectations of the educational technology system, and this variance expressed itself in multiple ways. Based on this, I used a few simple methods to test whether the topic model was identifying these essays in a systematic way.

To recognize essays responding to the prompt question by pasting it into the text box, as a structuring technique, I searched for exact phrases from the writing prompt and tagged essays if those lines appeared in the text. Essays were tagged as using the duplication tactic if any of these strings matched exactly in the paragraph.

Next, the `Skeptic` category is intrinsically oppositional to the system, but uses well-formed text to write in response to the system itself. I use a heuristic to identify essays in this mode by searching for reference to the 350-word minimum, which is a useful proxy for

metalanguage about the DAACS task itself:

> *"How much longer must I type? It seems as though this has to be at least 350 words at this point, but I will copy and paste this into word to check. Can you believe I have only typed 267 words as of the last sentence? We are getting close but I must draw this out in order to pass the assessment, foolish right? I am wondering if anyone actually reads these or if it is graded by a computer. My guess would be it is computerized as it would be as time consuming to grade these as it would be to type them. I am a huge advocate of efficiency so tasks such as this drive me crazy. Oh look! 350 words!"*

Finally, many students wrote authentic text but fell short of the 350 word minimum required by the DAACS system to proceed. This behavior was identified by vocabulary extraction into types (number of unique words) and tokens (number of total words). Essay type/token ratio was calculated to determine how much unique text actually appeared in a student essay; a lower ratio implies fewer unique words given the length of the essay, and below a threshold of 3:1, a corpus inspection indicated 100% precision at identifying essays with duplicated text.

Topic	Subtopic	Pasting	Minimum Reference	Type-Token
Structure	Introduction	2.7	0	2.28
Structure	Excitement	0	0.1	2.25
Structure	Commitment	0	0.2	2.23
Strategies	5PE	0	0.2	2.32
Strategies	Help-Seeking	0	0.4	2.26
Strategies	Environment	0	0	2.25
Strategies	Time	0	0	2.21
Motivation	5PE	0	0.1	2.29
Motivation	Anxiety	0	0.2	2.27
Motivation	Mindset	0	1.2	2.24
Motivation	Goals	0.1	1.2	2.19
Metacognition	5PE	0.0	0	2.28
Metacognition	Monitoring	0	0.1	2.28
Metacognition	Planning	0	0	2.27
Narrative	Past	0	0	2.20
Narrative	Future	0	0	2.29
Non-adherent	Pasting	4.4	6.3	2.52
Non-adherent	Skeptic	0.2	0.8	2.21

Table 34: Percent of paragraphs containing exact pasted text from DAACS interface and reference to word count minimum, by topic; type-token ratio of *documents* containing each topic.

My results are shown in Table 34. The table shows that both heuristic behaviors from the text occur disproportionately in non-adherent topics, and that documents labeled with the `Pasting` topic in particular have the highest type-token ratio of all topics. Notably, the introduction topic is the next most likely topic to be assigned to

paragraphs with exact excerpts from DAACS, as many students use those lines as part of their essay in otherwise good-faith paragraphs.

Sensitivity to Number of Topics

All of the results above are presented on a learned model with 20 unsupervised topics. But LDA is sensitive to the number of topics learned - the results may not generalize when the specific number changes. To test this, I conducted a reanalysis on different model sizes, with a step size of 4, creating new LDA models with 4, 8, 12, and 16 topics to compare to the model above. I then generate a hierarchy of topics using two different methods, one that produces a tree and one that produces a directed graph.

In the first approach, the *hierarchical* method, I take the following steps:

1. For each learned model starting at k=4 topics, I generate the paragraph-level distribution of topics and assign paragraphs a value for each topic based on those distributions.
2. For each topic in a learned model, I calculate the correlation coefficient between that model's values for each paragraph and each topic in the prior, coarser learned LDA model.
3. Beginning with k=8 topics, I assign a topic's "parent" in the prior model as the topic with the highest correlation coefficient across paragraphs.

Following this method I generate a hierarchy with 60 total topics in five layers; each layer represents a splitting into 4 additional topics. This method is similar in concept to hierarchical clustering, but allows each layer to be generated independently and only post-hoc aligned to topics in the previous layer. However, one downside of this approach is that it makes the assumption that topics group together over time, like constituents in a tree. This is not borne out in practice - in some cases, topics appear and reappear in subsequent runs on the same data, especially at different granularities.

So my second approach, the *percent overlap* method applies the following alternate method: For each learned model starting at k=4 topics, I assign paragraphs a label for each topic as before. Then, for each topic from the original 20-topic model, I measure the percent overlap of paragraphs that were also assigned each topic in the less granular models. This approach allows me to "anchor" at the original model and then study the stability of those topics as the granularity decreases. The advantage of this is that it acknowledges the non-hierarchical nature of topic modeling. One downside is that it is

harder to visualize; another is that the analysis is only suitable for evaluating all other topics against a reference topic model, in this case the original 20-topic model, and is no longer agnostic to the "correct" granularity.

My goal in the analysis that follows will be to evaluate the stability of topic hierarchies that appear in *both* of the methods described, to attempt to define a coherent and convergent story that is stable as topic count changes.

Results

The results of the hierarchical method are visualized in Figure 39; the agglomeration means that a tree is formed and can be drawn in a straightforward fashion. The non-constituency of the overlap method results in the topic assignments in Table 35. In both cases, the specific names that are used for topics in less-granular tiers are not as important as the grouping of topics from the $k = 20$ model and the consistency from one model to the next.

From this data, several patterns are immediately identifiable for their consistency.

- Time management and environment management are grouped together for simpler models but consistently separated from all other topics all the way back to $k = 8$ in both models; they are only separated at $k = 20$.
- Motivation topics are grouped together very early and only diverge into separate topics over time. Topics related to Anxiety diverge early from this group and maintain a stable topic separate from other Motivation topics beginning at $k = 12$.
- Help-seeking topics, though it references material from the Strategies section of DAACS, is more closely aligned with metacognition topics like planning and evaluation.
- Introduction paragraphs are identifiable very early as a unique topic, but use such formulaic language that they overlap with the non-adherent pasting of DAACS interface text all the way through the model at $k = 16$.
- Non-adherent skeptic language is grouped in with metacognition and help-seeking behavior in both analyses.
- Narrative language, whether past- or future-oriented, is separated out from the topics related to body paragraph subsections. Past-oriented backstory narratives are more closely associated with topics related to conclusion paragraphs, in particular, in both analyses of the topic hierarchies.

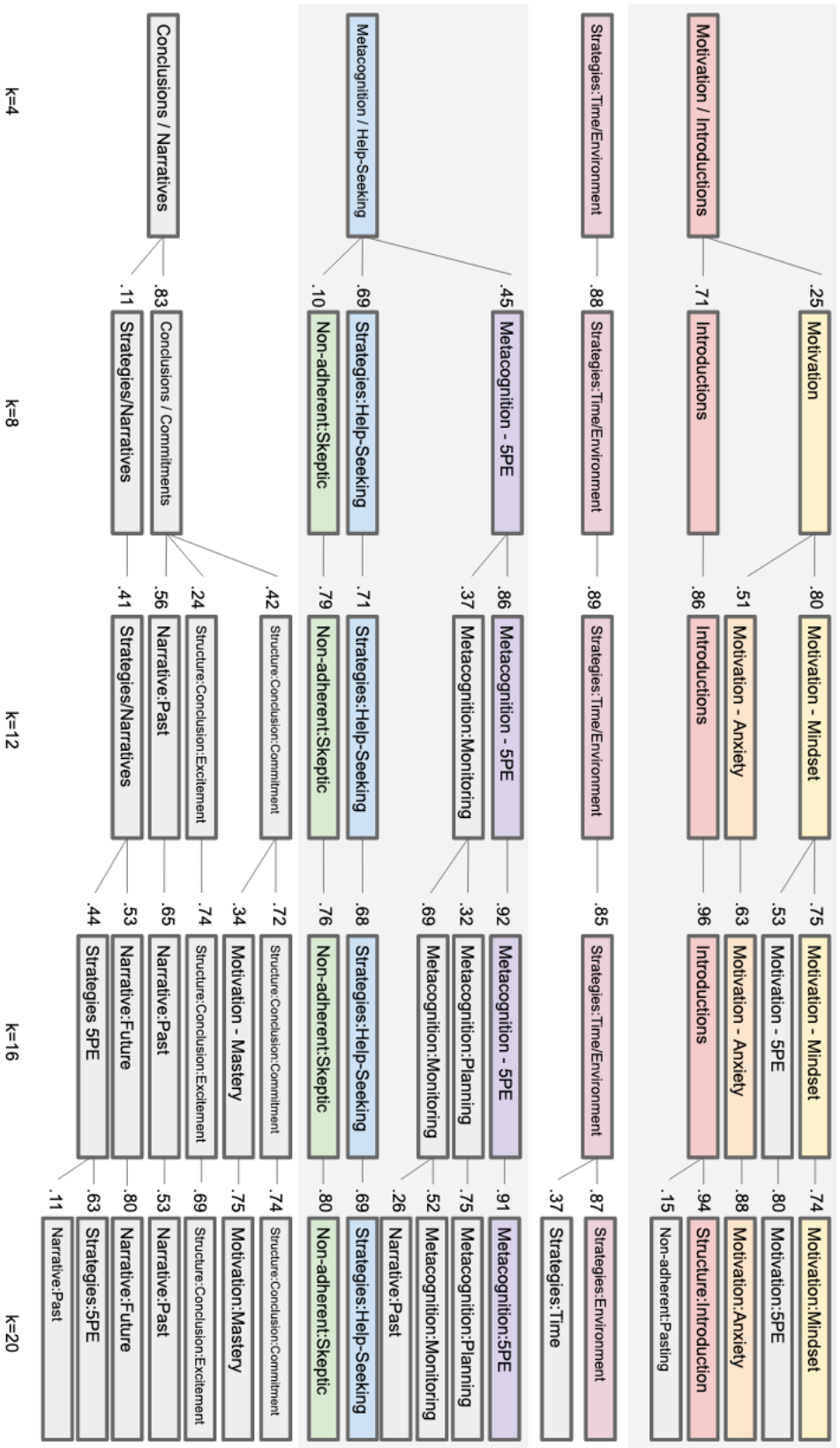


Figure 39: Relationship between topics as number of topics increases from 4 to 20, following the hierarchical method. Values between cells indicate correlation coefficient between topics. Topics with stable relationships over time are highlighted.

- The overview "5PE" topics are grouped together at the $k = 4$ level in the overlap analysis only; I label this topic the "Form Language" topic in this analysis. Because of the tree shape constraints in the hierarchical analysis, this grouping is not discoerable there.

	$k = 4$		$k = 8$		$k = 12$		$k = 16$	$k = 20$
61	Form Language	64	Introductions	55	Introductions	49	Introductions	Non-adherent:Pasting
93	Form Language	75	Introductions	91	Introductions	90	Introductions	Structure:Introduction
51	Form Language	37	Introductions	35	Narrative:Future	56	Strategies:5PE	Strategies:5PE
53	Form Language	89	Metacognition	86	Metacognition	89	Monitoring	Metacognition:5PE
90	Form Language	85	Motivation	68	Mindset	63	Motivation:5PE	Motivation:5PE
64	Form Language	79	Motivation	78	Mindset	79	Mindset	Motivation:Mindset
56	Narrative	29	Motivation	58	Goals/Commitment	59	Motivation:Goals	Motivation:Goals
40	Narrative	32	Narrative:Future	69	Excitement	57	Excitement	Conclusion:Excitement
78	Narrative	59	Narrative:Past	74	Goals/Commitment	71	Commitment	Conclusion:Commitment
63	Narrative	48	Narrative:Past	24	Narrative:Past	23	Narrative:Past	Narrative:Past
58	Metacognition	70	Help-Seeking	56	Help-Seeking	66	Help-Seeking	Strategies:Help-Seeking
92	Metacognition	50	Help-Seeking	32	Help-Seeking	25	Monitoring	Metacognition:Monitoring
46	Metacognition	33	Metacognition	24	Planning	67	Planning	Metacognition:Planning
47	Metacognition	66	Skepticism	59	Skepticism	68	Skepticism	Non-adherent:Skeptic
95	Strategies	95	Time/Environment	75	Time/Environment	90	Time/Environment	Strategies:Environment
74	Strategies	77	Time/Environment	65	Time/Environment	52	Time/Environment	Strategies:Time
59	Strategies	47	Motivation	57	Anxiety	82	Anxiety	Motivation:Anxiety
50	Strategies	72	Narrative:Future	58	Narrative:Future	68	Narrative:Future	Narrative:Future

Table 35: Relationship between topics as number of topics increases from 4 to 20 using the percent-overlap method. Values in cells for $k = 4 - 16$ represent overlap in paragraphs compared to the topic at $k = 20$ in each row.

I conclude that the most consistent topics, those which are least susceptible to variation based on topic granularity, are those related to anxiety, time and environment management, introduction paragraphs. These strongest signals also tend to be those that had significant relationships with scores and demographics in the prior analyses, suggesting that the cleanly identifiable topics are, in general, those that have significant external relationships with other factors as well.

Future Directions

My work on Wikipedia was, for the most part, hypothetical. No automated system for deletion decisions is in place; the automation is largely explanatory for understanding a domain. The ideas for decision support tools are prospective and aspirational. Using these explanations is somewhat thornier for automated essay scoring, though, because the system is *live* as a learning analytics tool today, and directly interacting with students in real-world environments; the stakes are raised.

Improvements in Understanding Students

Researchers in natural language processing always claim to be interested in greater insight into the datasets they work on. Better identification of errors based on genuine understanding should lead to performance gains in reliability, rather than blind improvement by an algorithm on the dataset as a whole. But finding out what those categories *are* is a technically difficult challenge. Clustering methods, either using LDA as I did here, or using more modern neural methods in embedding spaces[411] (which accomplish similar things), nevertheless requires a great deal of human insight and judgment to make sense. So I might ask next what it would take to discover these differences between subsets of the corpus *automatically*. While my work here looked at a few, specific things that students might do, like follow a five-paragraph essay structure, there is room for much more expansion. An expansion into more sophisticated automated topic discovery might also lead to discovery of personas that students take on in their writing[412]; this work might also extend to insight into how students express their own biography[413], a crucial step for understanding the student responses that do *not* follow the five-paragraph essay form.

This is important for the prospect of genuine personalized feedback that does not simply shunt students to a preordained writing form. Formative AES tools make claims of supporting student agency and growth; here, adapting to writer individuality is a major current gap. But recent commentary by Dixon-Román raises a host of questions about these topics specifically in the context of AES, asking how algorithmic intervention can produce strong writers rather than merely good essays. The critique, specifically, argues that:

[411] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453

[412] David Bamman, Brendan O'Connor, and Noah A Smith. "Learning latent personas of film characters". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 352–361

[413] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. "Gender identity and lexical variation in social media". In: *Journal of Sociolinguistics* 18.2 (2014), pp. 135–160

> *"revision, as adjudicated by the platform, is [...] a re-direction toward the predetermined shape of the ideal written form [...] a puzzle-doer recursively consulting the image on the puzzle-box, not that of author returning to their words to make them more lucid, descriptive, or forceful."*

This critique is valid: research on machine translation, for instance, has shown that writer style is not preserved across languages when an algorithmic system intervenes[414]. So for AES to adapt to individual writer styles and give feedback based on *individual* writing rather than exemplars of a particular form is uncharted territory. Natural language understanding researchers now argue that *"...style is formed by a complex combination of different stylistic factors"*[415]; Style-specific natural language generation has shown promise in other domains[416,417] and has been extended not just to individual preferences but also to overlapping identities based on attitudes like sentiment and personal attributes like gender[418]. For assessment, "authorial voice" has measurable outcomes on writing outcomes[419], while individual expression is central to decades of pedagogy[420]. Moving the field toward individual expression and away from those preordained forms may be a path to lending legitimacy to AES.

I might also suggest that the field move on to build specific *tools* that would make use of all of these findings on genre norms, demographic skews of topics, and non-adherence; but this would receive pushback in practice. In composition studies, concerns over the implementation of AES have largely been pedagogical rather than driven by any empirical insight about the underlying models or what they learn. This is not the approach that is taken in most other digital learning tools. Writing program administrators have written detailed, thoughtful practical guides on tools like e-Portfolios and digital instruction[421]. A burgeoning field of writing analytics is beginning to develop[422]. Even similarity checkers and anti-plagiarism software has been engaged with thoughtfully by composition scholars[423]. Yet a "discourse of rejection" has prevented similar engagement with AES development from many composition scholars[424].

[414] Ella Rabinovich et al. "Personalized Machine Translation: Preserving Original Author Traits". In: *Proceedings of the European Chapter of the Association for Computational Linguistics*. 2017, pp. 1074–1084

[415] Dongyeop Kang and Eduard Hovy. "xSLUE: A Benchmark and Analysis Platform for Cross-Style Language Understanding and Evaluation". In: *arXiv preprint arXiv:1911.03663* (2019)

[416] **hu17**

[417] Shrimai Prabhumoye et al. "Style Transfer Through Back-Translation". In: *Proceedings of the Association for Computational Linguistics*. 2018, pp. 866–876

[418] Sandeep Subramanian et al. "Multiple-attribute text style transfer". In: *Age* 18.24 (), p. 65

[419] Paul Kei Matsuda and Christine M Tardy. "Voice in academic writing: The rhetorical construction of author identity in blind manuscript review". In: *English for Specific Purposes* 26.2 (2007), pp. 235–249

[420] Peter Elbow. "Closing my eyes as I speak: An argument for ignoring audience". In: *College English* 49.1 (1987), pp. 50–69

[421] Edward M White, Norbert Elliot, and Irvin Peckham. *Very like a whale: The assessment of writing programs*. University Press of Colorado, 2015

[422] Joe Moxley et al. "Writing analytics: Conceptualization of a multidisciplinary field". In: *Journal of Writing Analytics* 1 (2017)

[423] Sandra Jamieson. "Is it plagiarism or patchwriting? Toward a nuanced definition". In: *Handbook of academic integrity* (2016), pp. 503–518

[424] Carl Whithaus. "Always already: Automated essay scoring and grammar checkers in college writing courses". In: *Machine scoring of student essays: Truth and consequences* (2006), pp. 166–176

Despite this, there is a growing evidence base that AES has a narrow role as a useful pedagogical tool[425,426]. Even critical pedagogy scholars, long the most skeptical of technological change, have recommended an integrated approach that acknowledges the value of new technologies when used alongside an empowering curriculum centered on student voices and needs[427]. The opportunity is open, but an understanding of the policy landscape is crucial for technology developers, lest they miss the broader cultural space that they are attempting to intervene in.

[425] Elena Cotos. "Automated Writing Analysis for writing pedagogy: From healthy tension to tangible prospects". In: *Writing and Pedagogy* 6 (2015), p. 1

[426] Alex Helberg et al. "Teaching textual awareness with DocuScope: Using corpus-driven tools and reflection to support students' written decision-making". In: *Assessing Writing* 38 (2018), pp. 40–45

[427] Nadia Behizadeh. "Realizing powerful writing pedagogy in US public schools". In: *Pedagogies: An International Journal* 14.4 (2019), pp. 261–279

Education Policy

In that bigger picture, my data is a jumping-off point for the decision-making of writing centers, first-year writing instructors, and teachers broadly. My results are especially striking for the differences they showcase between White women and other students as part of their introduction into college writing. Across a variety of analyses, my work showed that those women are culturally prepared for performing the compositional acts that receive high scores from both human raters and the automated systems that learn from them; college preparedness here may simply mean signaling membership in the genre norms that are taught at privileged high schools.

Yet this is not necessarily a sign of intrinsic skill. *"The genders are more alike than they are different,"* in writing assessment more broadly[428], and student skill measurement through grading is subjective and negotiated between students, instructors, and the school environment that they are in. The current consensus of the research community is that while cognitive factors play some part in gender differences in writing, the stronger effect is often tied to *attitude* and *motivation*, rather than any strict biological or developmental difference[429]. Assessment is a *"a social technique which has social consequences."*[430]; environmental factors, like parental education level and personal preferences of instructors, have a stronger effect on students' measured writing ability compared to the effect of innate characteristics like biological sex. Modern scholarship now recognizes the performative elements of racial language variation, including code-switching between dialects, as a discursive and performative practice for signaling prestige, group membership, and other social factors[431].

My focus on adherence, meanwhile, is not just about knowledge of genre norms, but also about permission to abuse a system or refuse it entirely. The issue is not one of *cheating*, *per se*, but of license and permission to refuse compliance with an automated system – permission that is potentially linked to identity. In automated, interactive systems elsewhere, abuse or non-compliance can make up from 10%

[428] Janet Shibley Hyde. "The gender similarities hypothesis." In: *American psychologist* 60.6 (2005), p. 581

[429] Diana Raufelder, Sandra Scherber, and Megan A Wood. "The interplay between adolescents' perceptions of teacher-student relationships and their academic self-regulation: Does liking a specific teacher matter?" In: *Psychology in the Schools* 53.7 (2016), pp. 736–750

[430] Barbara Read, Becky Francis, and Jocelyn Robson. "Gender,'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays". In: *Assessment & Evaluation in Higher Education* 30.3 (2005), pp. 241–260

[431] Ramon Antonio Martinez. *Spanglish is spoken here: Making sense ofSpanish-English code-switching and language ideologies in a sixth-grade English language arts classroom*. University of California, Los Angeles, 2009

to almost half of interactions[432]. In schools, there exist informal or implicit "negotiated understandings" about how students should conduct themselves and what rules they are allowed to break[433]. In this context of studying non-adherence, one thread to unravel in my work is the evidence that this skeptical behavior appears among men, and White men specifically.

This challenges or at least gives texture to a broader, older consensus that acting in defiance to school settings is a characteristic of marginalized groups, specifically Black students[434]. Yes, these understandings can often be implicit and coded, they can create barriers to community participation; for instance, on Stack Overflow, fear of hostile feedback for improperly meeting expectations of information seekers can prevent new users from asking questions or joining the community in the first place[435]. But it is hard to take my data and align it to the thesis of older work, that non-adherent student behavior is a feature of marginalized cultures. Instead, all of the data I found on commitment narratives and personal backstory suggests a highly motivated school culture among the POC students in my data. This supports the more contemporary research hypothesizing a more important role for self-determination than oppositional defiance[436].

Given all this, we should take seriously the findings in my work on the differences in what students write about and how they focus their essays. I've found that students vary both structurally and in what choices they make on what to write about. This should frame the future of research on AES in real-world, deployed contexts, where students are not simply supplying us their text for the sake of measurement of a skill; they are expressing themselves in their writing, and that writing hints at the broader systems those students are writing within.

For educators, this investigation boils down to a study of how students engage with academic culture, both in and out of the classroom. I don't take a specific stance in this dissertation on whether AES systems are an appropriate fit for any particular college campus or student population; in the absence of a complete failure or overwhelming evidence of fairness issues in a particular dataset and trained model, contextual factors for each school remain the most important deciding factor. In western countries, affluent families from well-resourced suburbs teach a rubric-friendly way of writing[437]. While the five-paragraph form is limited, it establishes a baseline structural style that scores well on DAACS, and this scoring trend is reproduced and even extended with automation. This is an important finding, and having this quantitative set of evidence should inform decision-making in writing program administration. While self-disclosure in writing is not necessarily a better or worse sign

[432] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. "Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse". In: *Proceedings of CHI*. 2020, pp. 1–13

[433] Neil Selwyn. "Exploring the 'digital disconnect'between net-savvy students and their schools". In: *Learning, Media and Technology* 31.1 (2006), pp. 5–17

[434] Signithia Fordham and John U Ogbu. "Black students' school success: Coping with the "burden of 'acting white"'". In: *The urban review* 18.3 (1986), pp. 176–206

[435] Denae Ford et al. "Paradise unplugged: Identifying barriers for female participation on stack overflow". In: *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM. 2016, pp. 846–857

[436] Kevin Cokley. "What do we know about the motivation of African American students? Challenging the" anti-intellectual" myth". In: *Harvard educational review* 73.4 (2003), pp. 524–558

[437] Martin Carnoy and Emma Garcia. "Five Key Trends in US Student Performance: Progress by Blacks and Hispanics, the Takeoff of Asians, the Stall of Non-English Speakers, the Persistence of Socioeconomic Gaps, and the Damaging Effect of Highly Segregated Schools." In: *Economic Policy Institute* (2017)

of strong writing, the students who choose to write about it receive different scores than their peers. Personal narratives and backstories appear in text differently than five-paragraph, topic-focused paragraphs. All of this is important to know about and will only get clearer as more advanced technical methods are developed.

My hope in all of this is that better, socially aware explanations will give composition scholars a handhold, some way to grapple with automated essay scoring systems in an open dialogue with technical researchers. In so doing, they may have a chance to use those tools to shape their own work, in a positive way. But they may also find that the explanations are a component of a better dialogue on-campus, and produce a more informed, more accessible advocacy around how to teach and measure writing.

Part IV: Takeaways

I've now shown that a rich, domain-specific and non-causal approach to explaining model behavior is effective for illuminating two domains: group decision-making debates and writing assessment. My goal now is to generalize, to describe a conceptual model for how technical researchers should give domain experts plenty to chew on and understand. The goal is to produce trust and confidence, rather than leaving users feeling vulnerable in the hands of a black box machine learning algorithm.

I next build up an abstraction for how we should think about model explanations. Leaning on a non-causal account of scientific explanation, I argue that the work I've done in this thesis shows how we can build justified stories about our models. Based on my work with Diyi Yang[438], I show one way to manage the needs of these explanations using a framework built around actions, intentions, goals, and circumstances. I tie this together with practical recommendations not just for machine learning researchers but the non-technical partners that help build algorithmic decision-making.

In so doing, I find that the context in which an algorithmic system is built matters. Part of this context is the purpose of the system itself, and our ultimate goal is to build systems that make the *right* decisions. Many of the applications that we develop in natural language processing and machine learning fill fundamentally problematic roles in the education ecosystem. For better or worse, in building explanations based on social and cultural context, the explanation ends up highlighting power hierarchies and relationships that the algorithmic system is designed to reinforce. And so I end this dissertation by confronting this tension. Based on my work with Michael Madaio and colleagues early this year[439], I leave with suggestions on how to build systems where their actions are not only explainable but just.

[438] Diyi Yang et al. "Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities". In: *Proceedings of CHI*. ACM. 2019, p. 344

[439] Michael Madaio et al. "Confronting Inherent Inequities in AI for Education". In: *International Journal of Artificial Intelligence in Education* (Under review)

Defensible Explanations

We have now learned a lot about our two domains. In both, the explanations for the human behaviors in the data came from a sufficiently reliable model; in neither case did I need to dive deep into model internals or specific weights in order to make discoveries about the population under study. Instead I looked at the outward behavior of the predictions, and investigated specific topics based on ties into the domain literature. But what commonalities are there between these two parts of the research? Perhaps a better question is: what was unique about these approaches to understanding my domains? Which of our insights could not come from a narrow, model-introspective, causal explanation strategy? And can we generalize this to best practices and a theoretical foundation for researchers working in new domains?

To answer these questions, let's go back to the philosophy of science. We have shown already that causal explanations fail for machine learning, but this new approach, acknowledging the social nature of people, their interaction with computational agents, and algorithmic decision-making, is something new altogether.

Return to Philosophy of Explanation

In their work on non-causal explanations, Robert Batterman and Collin Rice ask:

> *"How can a model that really looks nothing like any system it is supposed to "represent" play a role in allowing us to understand and explain the behavior of that system?"* [440]

[440] Robert W Batterman and Collin C Rice. "Minimal model explanations". In: *Philosophy of Science* 81.3 (2014), pp. 349–376

This question is important and relevant for our approach to explaining phenomena and learning about both of our target domains. Editors on Wikipedia argue the importance of their debate discourse and the reasoning and problem solving that it produces. Are they wrong to push back against automated tools that could replicate that judgment? Even more fiercely, educators will push back against automated scoring of essays. What does it mean to receive a score for an essay that was never read? These feel alien to us - and so any explanatory value that we can gain from them must be strictly scrutinized, especially if it does not come from the standard scientific playbook of intervention, causal relationships, and controlled trials.

Minimal Models

Batterman and Rice, who were quoted above, have a suggestion for us. Their goal is to give an account of models that are explanatory, giving new insight and scientific knowledge, without causal mechanisms or interventions. They argue that as scientists we do this by isolating recurring, co-occurring features of a phenomenon, and connecting them using analogies. Many accounts of explanation only allow us to do this if we have a specific and accurate understanding of the exact relational links between entities. Their minimal model account, on the other hand, describes the scientific process in cases where we cannot isolate the exact relations, and have no way to tease them apart. In these cases, the specifics of causal mechanisms are not always important and the stories that result are not always perfectly representative or factual. But non-causal accounts argues that a model meeting "extremely minimal" accuracy conditions, conditions that causal explanations would find "not terribly important", can still give a successful explanation.

Consider two examples from hard sciences: population biology and fluid dynamics. In naturally occurring settings, we cannot give full causal stories, only caricatures. We can list off the features that are relevant in those populations – fitness functions, optimizations, tendencies of variables to correlate and phenomena to co-occur, but we can't draw causal links, again because things are so entwined. If we relied on pure causal models in these settings, we couldn't make sense of the world. The evidence we have is simply "insufficient for deriving the target explanandum." Yet in both settings, scientists know a set of factors that must be in place in order for phenomena that we see happening in real life to occur, like an equal ratio of male to female individuals in a species population. But the specific factors are all tangled up – it "stretches the imagination" to think of any one observed phenomenon as a causal factor. Instead, there are merely common features, co-occurring.

Despite natural scientists' models being "minimally accurate" they are robust and explanatory! Natural scientists, in their non-causal research, learn something important about how the world might actually work[441]. The work involved in justifying these models involves iteratively discovering certain features of a population, determining whether those features are relevant, or whether they can be ignored.

Scientific models in these circumstances take the first step toward being explanatory by idealizing away the factors that are not relevant to the final outcome, labeling them irrelevant. In the minimal model account, scientists tell a story of why large classes of *other* features are *irrelevant* to the explanation. Philosophers argue that non-causal

[441] Jon Seger and JW Stubblefield. "Theoretical Evolutionary Ecology". In: *Bulletin of Mathematical Biology* 4.58 (1996), pp. 813–814

explanations answer those questions by constructing a space of possible systems, defining the boundaries of a class of systems that our factors *do* explain. The power of the explanation comes from being able to clearly delimit the scope of the types of systems that fit our explanations.

But we need to then answer further questions. Can we determine the class of systems that will follow these behaviors? How stable are the behaviors under certain changes in the system? To answer these questions, science needs to tell a story of which populations will exhibit the tendencies described – populations where the set of phenomena, together, are subject to the same constraints. This involves defining what entities are being modeled, what interactions between those entities are being described, and the context in which those interactions occurred. The resulting model is as minimal as we can define it, producing a certain "universality class," common circumstances in which our explanation and features apply. From there, we can use the knowledge we gained: If we make a claim about a class of entangled circumstances, and define how to test whether a new set of data comes from that class, then claims made about the class as a whole can tell a story about any constituent member of the class. We then argue that other details – including details we get wrong! – are less relevant, because they aren't needed to define the boundaries, and do not prevent the defining characteristics of the class from generating new insight.

These models, and the lessons learned from them, explain how heterogenous systems end up adopting similar strategies or following similar patterns. The hard task left before us is to justify what the constraints of our universality class *are*. How do we get that justification?

Structural Models

The work of Alisa Bokulich is useful for understanding the justification process behind a non-causal model. Working in parallel to the acount above, she also argues that the strict, Woodward-style interventionist research makes unsustainable claims.

> *[causal-only explanation] "...would suggest that scientists rarely - if ever - succeed in offering explanations – even when there is a consensus in the scientific community that an adequate explanation has been given."* [442]

[442] Alisa Bokulich. "Distinguishing explanatory from nonexplanatory fictions". In: *Philosophy of Science* 79.5 (2012), pp. 725–737

This account of explanation presents a similar non-causal, model-based account of explanation to the proposal by Batterman and Rice, above. But it goes a little further: in this alternate view of non-causal explanation, it is not just the "true parts" that do the explaining, but

the *fictions* as well. Models meeting these criteria are not "merely phenomenological models, useful tools for making predictions." Instead, they are capable of "generating real knowledge and genuine insight." But in order for a non-causal explanatory model to be successful, Bokulich argues it must meet three criteria:

- The explanans makes reference to a *scientific model*, which is to some degree idealized or fictitious.
- The model shows how the elements correctly capture patterns of counterfactual dependence - they “reproduce" the relevant features.
- A “justificatory step" specifies the domain of applicability of the model, and the extent to which the model can be trusted, for which purposes.

The first of these steps, in our case, refers to the trained classifier that attempts to accurately capture the decision-making of a particular domain. The second step here is captured by the features that I have been documenting throughout the last several chapters of this thesis - phenomena and insight about the domain that the model is meant to automate. But what remains is the third step. Bokulich's account of this step aligns to Batterman and Rice's idea of a universality class, but gives some additional instruction on how to actually make this argument. Bokulich specifically addresses the problem of being *overly* permissive of explanations. If *anyone* can tell a good story for why a set of variables occurring together are explanatory, and claim it extends to any related domain, then we are in danger. Relevance relations are asymmetric, and causality is *real*, even if it is sometimes hopelessly entangled in variables that cannot be surgically intervened on.

To answer this, Bokulich suggests one needs to turn to the nitty-gritty details of the science in question. When evaluating whether an explanation applies to a domain and is successful, we must openly acknowledge the current state of the scientific field, as part of the explanation itself. Given the knowledge that we had at the time the explanation was made, would the scientific community give the arguments credence? This approach allows us to acknowledge that scientific practice is fundamentally *discursive*, collectively built just as the Wikipedia debates we study in the data itself. What counts as an adequate explanatory representation is something that has to be negotiated by people, working together in a domain and aided by the features captured in the model.

Bokulich calls the process of building this context, in defense of an explanation that does not make causal claims, the *justificatory step*.

This step is the process of building up the universality class of an explanation. It is this justification that must do the heavy lifting, and it is fundamentally a *negotiation*, rather than a causal chain. In this negotiation, we must, as scientists, account for the circumstances of the study, the details of the domain, and the purposes for which the scientists are building and deploying a model. An explanation detached from these contexts is insufficient; but an explanation using non-causal evidence and entangled variables, contextualized with this evidence, scientific discourse, and limits on applicability, can be successful.

Negotiating a Justified Explanation

So Bokulich, Batterman, and Rice have given us a set of tools to make non-causal arguments. We must work from exclusion, defining the bounds of the universality class, circumstances where our explanations hold water. We must clarify explicitly that our model is minimal, establishing a set of patterns that, collectively, explain behaviors while claiming that other factors that make different cases distinct are irrelevant to the mechanisms that we've described as part of our explanation. And we must justify why the set of factors we've identified are relevant and important to the explanation, leaning on something other than the internals of our dataset for this justification.

This aligns with prior work in explainable machine learning. While other authors in this space have never specifically aligned their findings to the accounts from philosophy, they have come to similar conclusions about the importance of a justification that fits in the contemporary discourse. Deciding on and subsequently defending the factors of an algorithm is a subjective, "editorial" process[443], designed by human operators to automate human judgment. For users, introducing an unknown technology here is a "leap of faith," as the internal processes can't be observed or understood[444]. And so a suitable explanation must not be technically correct as much as it must be trusted.

A social approach to explainable machine learning supports the case for "why" questions, downplaying the importance of specific technical implementation details. Intriguingly, corporations seem to know this in their own attempts at explanation. Most information from Facebook, for instance targets the motivating "why" questions behind their algorithms rather than technical "how" questions[445]. But corporations attempt to tie trust in automated decision-making to the reputation of the firm itself, as well as emphasizing future usability and benefits, and it's unclear to whom that approach is satisfying or successful[446].

[443] Tarleton Gillespie. "The relevance of algorithms". In: *Media technologies: Essays on communication, materiality, and society* 167.2014 (2014), p. 167

[444] Kevin Anthony Hoff and Masooda Bashir. "Trust in automation: Integrating empirical evidence on factors that influence trust". In: *Human factors* 57.3 (2015), pp. 407–434

[445] Kelley Cotter, Janghee Cho, and Emilee Rader. "Explaining the news feed algorithm: An analysis of the" News Feed FYI" blog". In: *Proceedings of CHI Extended Abstracts*. 2017

[446] Monika Hengstler, Ellen Enkel, and Selina Duelli. "Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices". In: *Technological Forecasting and Social Change* 105 (2016)

A Framework for Establishing Boundaries

In my work with Diyi Yang, she used data-driven methods to separate user intent and explain observed phenomena around tenure in the context of online health communities[447]. This framework conceptualizes roles using five facets, reproduced here, reordered for relevance.

[447] Diyi Yang. "Computational Social Roles". PhD thesis. Carnegie Mellon University, 2019

1. **Person.** Attributes of individuals, like their age, race, or gender. In the context of any one interaction, these attributes are relatively static, particularly in how they are perceived by others in computer-mediated interactions[448].

[448] Charles G Hill et al. "Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways?" In: *Proceedings of CHI*. ACM. 2017, pp. 6658–6671

2. **Goal.** Individuals participating in decision-making processes are not doing so irrationally or neutrally; they have some intention during the interaction. Identifying and describing these goals allows us to better understand how they interact with each other, in a group context, and what criteria they use when they make a decision.

3. **Interaction.** While personal aspects of identity above are sometimes "innate" and sometimes mapped to biological or physical characteristics, most are performative[449]. This means that attributes of an individual's identity and how they approach decision-making can change over time and based on who they are interacting with. A key aspect of explanation is recognizing how otherwise-static attributes of decision-makers change depending on the target of a decision and the audience for the decision's outcome.

[449] Candace West and Don H Zimmerman. "Doing gender". In: *Gender & society* 1.2 (1987), pp. 125–151

4. **Expectation.** Individuals do not arrive at a decision-making discussion teleported from a vacuum; they have experience throughout their lives in similar contexts, for "how decision-making happens." This includes the particular domain of the decision, if they are frequent participants or have a history of performing a choice to be automated.

5. **Context.** The specific topic under review, the venue and timing of a decision, and the recent history of past decisions all may alter how other aspects above are expressed in a particular decision.

I suggest that we can understand our models using this framework, and use it to situate our findings from an explanatory investigation. We can use each of these framing questions to define a wall or boundary, a limitation for where we think our findings can apply as successful explanations. When a new dataset or domain is brought in, we must ask whether it fits all of the criteria above before

we can trust that our previous evidence holds. When it does not, in one or more dimensions, it will be our responsibility to test our assumptions. But instead of testing the causal factors and measuring from a strictly interventionist account, we must instead test whether the feature that differs is a member of the same class of decisions as those in which our original explanatory features were discovered. If it is not, we can excise it from our definition; if it is, then we cannot make forecasts or predictions of how that decision will be made – our explanations simply do not apply.

Using the Framework: Wikipedia

This thesis does not attempt to rigorously test the framework for use in explanation, but we can look at the areas of focus in the chapters so far and attempt to apply this model to them. Research using early versions of this framework in the context of Wikipedia editors has already shown promise, discovering the granular intentions of individual edits to more intelligently categorize editor actions[450,451]. It has been useful for understanding editor roles over time, as well, describing for instance the way women are more likely to take part in emotional labor roles that are necessary to maintain basic community functioning despite lower associated prestige[452].

Explaining behavior by focusing on user goals, in the case of Wikipedia, might largely mean studying their stance on a debate. Certainly that has been the focus of most sentiment analysis or stance classification work in NLP. We can assume that a user's goals for a particular page align with their vote, typically for `Delete` or `Keep`. But just like with our DAACS results, the goals of users may diverge from good-faith debate content. Early work on administrator promotion[453], for instance, showed that raw edit counts and basic forms of politeness were sufficient predictors of administrator promotion; as a result, users seeking promotion sometimes go out of their way to "pad their stats" with relatively minor edits and other easily measurable social moves[454]. These self-centered goals from individuals may be distinct from having strong opinions on the outcome of any one debate, and may be separated from a user's goals as defined by their stance, depending on the direction of the research. Just like in DAACS, these transgressive behaviors have rich opportunity for explaining how decisions are made in a real, rather than idealized, debate setting.

So these three attributes are relatively well-defined immediately — a user's personal attributes, stance, and the nominated article as topic of discussion. Using the framework above tells us where we need to fill in gaps in future work. For this context, this means finding

[450] Diyi Yang et al. "Who Did What: Editor Role Identification in Wikipedia." In: *ICWSM*. 2016, pp. 446–455

[451] Diyi Yang et al. "Identifying semantic edit intentions from revisions in wikipedia". In: *Proceedings of EMNLP*. 2017, pp. 2000–2010

[452] Amanda Menking and Ingrid Erickson. "The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia". In: *Proceedings of CHI*. ACM. 2015, pp. 207–210

[453] Moira Burke and Robert Kraut. "Mopping up: modeling wikipedia promotion decisions". In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM. 2008, pp. 27–36

[454] Katie Derthick et al. "Collaborative sensemaking during admin permission granting in Wikipedia". In: *International Conference on Online Communities and Social Computing*. Springer. 2011, pp. 100–109

out more about the boundaries of interaction and expectation. My findings on policy tie into expectation – users are citing the policies from the past as part of their expectation of what *ought* to happen. The work I've done in those chapters serves not only to make predictions about outcomes but to indicate what contexts those predictions ought to hold in. Explanation through policy citation has immediate value — and is validated in its aligned to prior work. Pavalanathan et al.[455] demonstrated, for instance, that citing policies on talk pages *does* influence editing behavior.

But work remains in the final category, interaction. A good explainable model would recognize that behaviors like quantity of posts, timeliness of posts, and reply patterns *between* users are all candidates for inclusion in an explanation. Narrowing down the set that has explanatory value is part of the work of justifying a minimal model of the *AfD* domain.

[455] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. "Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), p. 137

Using the Framework: AES

In my essay scoring explanation, we have fit our two categorical demographic variables into the person aspect of this framework. My results implicitly test disparate outcomes for White against non-white students, and men against women without accounting for transgender or non-binary identities. The data exists in DAACS to evaluate for other personal identities, like age and military status; however, I have not yet explored what insights we can learn from those explanations or how they can shape the bounds of our explanations.

Student goals and expectations are perhaps the most interesting aspect of my AES research. My study of content, in particular, highlighted that goals vary for students based on their willingness to adhere to the format and genre of the automated system. Intersecting with identity, I showed that these goals and expectations can be very different for men, and White men in particular. Their intention with DAACS was not necessarily to achieve the highest possible score but merely to meet the bare minimum; future attempts at explaining the behavior of AES systems must account for these differential goals from students as part of any system description.

Student expectations can be thought of in terms of genre norms, and whether students believe they understand the "right" way to answer an essay prompt. This expectation around genre norms was reified for the model, though, not by students or by the rubric but by the annotators themselves. Their scores, after all, served as labels for the model where these differences surfaced. My structure analysis, focused on the five-paragraph essay, showed that this is a substantial underpinning of the entire corpus, but is not explic-

itly encoded anywhere in the rubric; students are not told to write a "five-paragraph essay" but some do – predominantly White women – and are rewarded for it with higher scores. Explanations of scoring model behavior in the future must be able to give an account of how it interacts with those structural expectations and who is advantaged by the hidden expectation of the system.

My content analysis also showed that students are coming to DAACS from startlingly different contexts. Some students, predominantly White women again, are coming from home environments that are difficult to manage, and are willing to disclose substantial vulnerability about test anxiety. Other students have no such additional constraints on their education experience. Future exploration, especially in the younger population of the 2020 dataset, will have to explore whether this context actually changes the way students interact with and write for the automated system.

Verification of the Framework

There are many ways to verify that this framework is useful for building explanations that are defensible and trustworthy in a real-world context. This will be an important future direction.

First, of course, is the NLP standby - technical evaluation and quantitative measurement. In my work with Diyi Yang, for instance, we codified the aspects and associated features in this framework into a Gaussian mixture model and were able to automatically define whole roles for community members at the intersection of all of these features. This enabled a *measurable* level of coherence for the features themselves and how they co-occurred. But this focus on the technical, the first place that quantitative data scientists turn, may be too closely tailored to the needs of the data scientists themselves[4][5][6].

[4][5][6] Harmanpreet Kaur et al. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning". In: *Proceedings of CHI*. 2020, pp. 1–14

Also in that work, user studies were used as a tool to evaluate whether the roles that emerged in that technical model were interpretable and matched the intuitions of domain experts. This involved multi-stage interviews and mixed-methods interpretations that led to the labels for role types. This method generated subject matter expert buy-in, rather than focus on the narrow view of interpretability, and allowed domain knowledge from praxis to drive the presentation of results and explanation for the domain under study. Observation of real-world phenomena is another tool for evaluating whether the explanations arising from this framework are defensible. We might actually deploy the systems that we propose above, and see whether they interact in the ways that we expect.

And from a final, more theoretical perspective, further engagement with philosophy of science is also going to be valuable to allow us to

establish more formal bounds and expectations for what counts as a good explanation. This will clarify the role of a minimal model and a justificatory step, allowing the field to further advance a humanities-based understanding of whether our models are useful reflections of the domains they attempt to automate, and whether our explanations are succeeding at their goals.

Algorithms as Social Actors

But there's something funny about the account we're giving right now. In most accounts of explanation, the model we describe is simply that - a *model*. It is used for explanation and understanding, the progression of scientific knowledge; the model is a minimal form of the behavior we observe and it lets us learn more.

But in applied machine learning specifically, these models have a dual purpose. Not only are the models explanatory for the human processes we're learning about; they are also directly active in the process being observed. This is uncharted territory - a simulacrum of the process, joining in as a *co-participant* is unlikely to be a problem that comes up in other fields. So let's investigate what this means.

As in the philosophical account, the context, goals, and interaction style of an automated decision-making system also matter. Algorithmic decision support in groups produces better outcomes when they explicitly express vulnerability[457]. And social, proactive robots are more likeable and viewed as more productive[458]. But this does have the result of lowering the effect of automation: warning about uncertainty produces lower trust and more frequent manual interventions[459]. Users working with algorithmic systems do not trust all behaviors to be automated equally. Studies have been shown, for instance, that people trust robotic ability in specific tasks like scheduling and workflow automation over others[460]. This is *certainly* the case in education technology, as we've seen in the numerous debates in both of our target domains.

More than two decades ago, Reeves & Nass showed that people can "team up" with a computer when they believe their performance relies on the computer's performance. They argued that computer systems are treated as *social actors* in how users interact with them[461]. In this social setting, explainable machine learning has the opportunity to address either the "how" questions and "why" questions about an algorithm in use[462]. But people are not focused on the formal details of explainability so much as a definition based on *trust* – to the point that robots making erroneous explanations are sometimes *more* likable than accurate ones[463].

[457] Sarah Strohkorb Sebo et al. "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 2018

[458] Guy Hoffman and Cynthia Breazeal. "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 2007, pp. 1–8

[459] Tove Helldin et al. "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving". In: *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 2013, pp. 210–217

[460] Matthew C Gombolay et al. "Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams". In: *Autonomous Robots* 39.3 (2015)

[461] Byron Reeves and Clifford Nass. *How people treat computers, television, and new media like real people and places*. 1996

[462] Wolter Pieters. "Explanation and trust: what to tell the user in security and AI?". In: *Ethics and information technology* 13.1 (2011), pp. 53–64

[463] Nicole Mirnig et al. "To err is robot: How humans assess and act toward an erroneous social robot". In: *Frontiers in Robotics and AI* 4 (2017), p. 21

Let's look at a few specific factors about trust in automated decision-making[464]. The literature here tells us that increasing the trustworthiness of an algorithmic system is not always tied to accuracy or reliability. Instead, the most successful trustworthy systems rely on *calibration* between user expectations and performance. High initial trust in a system can decrease following interaction[465], especially after seeing algorithmic decision-making commit errors or behave in ways that do not match the norms of human decision-makers that are being emulated. In reality, people base their trust in systems on its actual behavior, their perception of the algorithm's accuracy, and how that perceived accuracy aligns to the system's self-stated accuracy[466]. Facilitating trust comes from setting correct expectations about what an automated system is going to do[467,468].

Still, there is a lot of theory-building work to do. For practical reasons, it's rare that we can actually be fully descriptive about the tradeoffs that went into the design of a system; we also can't be fully clear to all users about the expectations about a system[469]. Our primary pointer for future directions is that while transparency in functionality is important, it is not everything. Explanation that succeeds in building trust is largely based on calibrating user expectations to actual performance, rather than actually walking through the specifics of the decision-making process. That practical guidance will hopefully shape the work still to come in the field, providing a unified account of the machine learning classifier as explanatory scientific model, and that same classifier as social co-participant in the decision-making process itself.

Unifying these two views of the classifier, though, will take us one step further up from the data than we've been so far. By allowing ourselves to intervene in decision-making directly, we give license to our algorithmic systems – permission to alter the course of human judgment. Before we do that, we should take a long look at what team we are supporting, and ensure we're advancing the causes that we believe in – and with that, we can move on.

[464] Ella Glikson and Anita Williams Woolley. "Human trust in Artificial Intelligence: Review of empirical research". In: *Academy of Management Annals* ja (2020)

[465] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), p. 114

[466] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. "Understanding the effect of accuracy on trust in machine learning models". In: *Proceedings of CHI*. 2019, pp. 1–12

[467] Bo Xiao and Izak Benbasat. "E-commerce product recommendation agents: use, characteristics, and impact". In: *MIS quarterly* 31.1 (2007)

[468] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. "Toward establishing trust in adaptive agents". In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2008

[469] Mike Ananny and Kate Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". In: *New Media & Society* 20.3 (2018), pp. 973–989

Confronting Inequity

Early in this thesis, I made the point that explanation is about epistemology, not ethics. The tools that I developed in the previous chapter for thinking about explanation stay in that mindset, focusing on justification of an explanation as successful. But the approach that I take to describing successful non-causal explanation requires an understanding of disciplinary discourses. An explanation is only successful if your justification, as a scientist, is constrained to the circumstances where it fits based on what your contemporary scientific discourse believes.

This is one point where the line between explanation and ethics blurs. It's also a jumping-off point for a broader and more skeptical look at our scientific process of explanation. This is because the choice of who is a member of that "scientific discourse" is itself fraught with societal biases and preferences for some groups over others; this is true throughout natural language processing and is certainly true in education, meaning we should expect it in how we define contemporary thought in education technology, too.

Our field is now waking up to the impact of our work, but our notions of *power structures* as they relate to algorithmic systems are still maturing[470]. Given the rapid scramble to technology-based learning platforms in the last academic year due to the coronavirus pandemic, it is now more urgent than ever for the fields of algorithmic decision-making, natural language processing, and education technology – as well as those who interact with that research – to catch up to the broader discourse about equity and justice in the application of automation to schools. So let's lean on scholarship from critical pedagogy in this final chapter, to interrogate the question of *who* gets to decide what makes an explanation successful.

[470] Su Lin Blodgett et al. "Language (technology) is power: The need to be explicit about NLP harms". In: *Proceedings of the Annual Meeting of the Association for Computational Lingustics (ACL)*. 2020

Disciplinary Norms and Power

Technology and Good Intentions

Recently, critical theorist Neil Selwyn engaged with the question of automation technology, especially under the banner of "learning analytics," and how it fits into the ongoing work to improve equity in education[471]. He argued that this research puts the institution ahead of the individual, ignoring the broader social context of technology.

[471] Neil Selwyn. "What's the Problem with Learning Analytics?" In: *Journal of Learning Analytics* 6.3 (2019), pp. 11–19

Work in predictive modeling and decision-making views algorithms as a means of surveillance rather than support, leading to limits on free choice and expression. This line of thinking follows a common critique against what theorists call *techno-solutionism*[472]. Fundamentally, work in this style assumes that individuals and organizations can add well-intentioned technology to their existing processes and, in doing so, solve deep-seated, complex social issues.

[472] Evgeny Morozov. *To save everything, click here: The folly of technological solutionism*. Public Affairs, 2013

Later, Selwyn extended his argument with one further step[473]. Educational data mining is not apolitical, he argued, and cannot be; every system and artifact of that system, like a dataset or a trained model, has politics embedded. Even "social good" targets like learning analytics require you to make some normative decisions (like "staying enrolled in a university course is good"). Professional codes of conduct among scientists and engineers rarely make explicit these normative judgments about what projects should be doing, but his provocation was that the field must evolve learning analytics toward a *deliberate* politics of social justice. This might require a rebuilding of the whole field.

[473] Neil Selwyn. "Re-imagining 'Learning Analytics'… a case for starting again?" In: *The Internet and Higher Education* (2020), p. 100745

Crucial to this analysis is the "master's tools" analysis borrowing from Audre Lorde[474], drawing a distinction about whether algorithmic technologies are "reformist" or not (do they work within an existing system?). He takes as an example a transgender student with a disjointed home life, meaning they have disrupted attendance, parental engagement, and medical records. Because those fields have gaps, omissions, or blanks, by definition this will lead to reduced fidelity in predictions. These gaps are built-in to the very *database schema* of algorithmic interventions; this makes it structurally impossible, at a technical level, to address inequalities head-on. Even in this thesis, we've seen this in the limited way I was able to engage with race and gender in my analysis of DAACS.

[474] Audre Lorde. "The master's tools will never dismantle the master's house". In: *Sister outsider: Essays and speeches* 1 (1984), pp. 10–14

When I was in industry, I did not have the room in my schedule to ask those questions, nor did I have any incentives to make that room. After our product had been in the market for a few years, I started being exposed to more outside experts in education, like Selwyn and his peers. Rather than coming at problems like AES from a machine learning perspective, hoping to help educators from the outside, they were starting from the perspective of students, educators, and community members watching technology seep into their daily lives. Interrogating the culture of education has an extremely long history, of course, going back decades[475,476]. Researcher-activists have built rich culturally sustaining pedagogies to accomplish goals of social justice for students[477]. And yet in my original years of academic and industry work on AES I got only fleeting pointers to this work, and no formal training.

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

[476] Paulo Freire. *Pedagogy of the oppressed*. Bloomsbury publishing USA, 1970

[477] Gloria Ladson-Billings. "Toward a theory of culturally relevant pedagogy". In: *American educational research journal* 32.3 (1995), pp. 465–491

What I have found now is that the question at this point is not *whether* inequities exist in educational technology. The unequal treatment of students in today's educational settings is well-established. The broader field of algorithmic decision-making for education has been slow to respond to this, though. While I was working on automated essay scoring, we certainly felt the backlash to our work[478,479]. But as a developer, the reaction in the moment is to be immediately defensive of the system, arguing that machine learning only encodes preexisting biases, and does not create new ones or cause additional harm. It certainly did not feel right to be held accountable for the state of the world we found going in, embodied and crystallized in our data[480]. This is typical; with only a few exceptions, the fairness debate has not yet penetrated the mainstream discourse of learning analytics or predictive education technology[481].

[478] NCTE. *NCTE Position Statement on Machine Scoring*. https://bit.ly/3dQHaVY. Accessed 2020-06-30. 2013. URL: https://bit.ly/3dQHaVY

[479] John Warner. "The Ed Tech Garbage Hype Machine: Behind the Scenes". In: *Inside Higher Ed* (2014). Accessed 2019-09-24. URL: https://bit.ly/1w3NdwS

[480] Brent Daniel Mittelstadt et al. "The ethics of algorithms: Mapping the debate". In: *Big Data & Society* 3.2 (2016)

[481] Kenneth Holstein et al. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of CHI*. 2018

Funding for Algorithms in Education

Part of the reason is that confronting systemic inequities through action is *difficult*. Individual researchers may feel unable to make significant changes to their research agenda, much of which has been built up over years and is hard to pivot. Even if they were to change an individual project to prioritize equity, doing so may feel small in the grand scheme of the learning science landscape, unlikely to effect change at a larger scale while putting their research funding at risk. This challenge in making systemic change is precisely the issue. Identifying differential performance or group fairness metrics for individual systems or for particular subpopulations is insufficient if those systems are built on, and further reproduce, existing systems of oppression. Change requires collective impact rather than individual action, which in turn requires mapping the network of power and funding that produces that impact[482]. In this section, I look closer at the funding incentives and structures that have shaped the current state of the learning sciences.

[482] Brian D Christens and Paula Tran Inzeo. "Widening the view: situating collective impact among frameworks for community-led change". In: *Community Development* 46.4 (2015), pp. 420–435

Funding for education research follows much the same cycle as other academic research in general. Calls for proposals are put out annually, workshops are funded, conferences are held and shared tasks are developed for collaboration across institutions. Funding is allocated to individual PIs and teams at the scale of hundreds of thousands of dollars at a time. But by their very nature, these grants determine rigid boundaries as to which forms of learning science they can support, and which evidence counts as success, providing further examples of bell hooks's argument that our "ways of knowing" – the types of research that drive the field forward – are circumscribed by historical relations of power[483]. In public guidance

[483] bell hooks bell. *Teaching community: A pedagogy of hope*. Vol. 36. Psychology Press, 2003

to grantwriters, a tenured professor and reviewer for IES makes clear which types of educational research are valued: *"If you can't provide a reasonable estimate of what your minimum sample size will be, and its power, then you are already dead in the water."*[484]

A successful application for later-stage grants in the funding pipeline, meanwhile, requires demonstrated outcomes from a pre-existing intervention at an earlier stage; the new work must be made up of relatively small, incremental changes to that existing result[485]. Progressing to the next tier of funding requires evidence that the previous step worked. Radical changes at any later point of a research program are difficult to justify—and indeed, radical re-visioning of the fundamental aims and goals of one's research line is made difficult by these processes.

The publication cycle for machine learning and learning sciences also reinforces this structure of iterative change to existing interventions. Performing small studies with clearly-defined alterations to prior work lends itself well to designing experimental conditions and results that fit within the bounds of conference submissions. This work produces discrete and clean findings that can answer narrow questions rigorously; such work can also be cited in time for submission of the next cycle of grant proposals. This structure means that researchers are strongly incentivized to maintain and build upon existing systems and make isolated changes to a deployed platform for experimentation — as in widely-used platforms like ASSISTMents[486], Betty's Brain[487], or Cognitive Tutor[488]. In this light, it is understandable why the bulk of work on fairness and ethics in educational technology and the machine learning field more broadly seems intent on thinking of, and operationalizing, fairness as a technical evaluation of models and datasets, rather than as systemic problems. Opportunities for a drastic rethinking of the goals of those research lines and educational platforms are rare and risky. Confronting inherent inequities requires a leap of faith for a PI – altering course and hoping that the funding is there to catch you when you jump.

[484] Stephen Porter. *Applying for an IES Grant*. 2015. URL: https://stephenporter.org/research-methods/applying-for-an-ies-grant/

[485] National Center for Special Education Research. *Building Evidence: What Comes After an Efficacy Study?* 2016

[486] Neil T Heffernan and Cristina Lindquist Heffernan. "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching". In: *International Journal of Artificial Intelligence in Education* 24.4 (2014), pp. 470–497

[487] Krittaya Leelawong and Gautam Biswas. "Designing learning by teaching agents: The Betty's Brain system". In: *International Journal of Artificial Intelligence in Education* 18.3 (2008), pp. 181–208

[488] Steven Ritter et al. "Cognitive Tutor: Applied research in mathematics education". In: *Psychonomic bulletin & review* 14.2 (2007), pp. 249–255

Conditions for Scientific Change

The Structure of Scientific Practice

To understand this cycle from a theoretical foundation, I would draw on Thomas Kuhn's paradigm for understanding scientific practice[489]. Scientific study, in his model, acts as a feedback loop. The norms and values of a particular scientific paradigm shape what is considered to be worthwhile science. Then, opportunities for funding, reflecting those paradigmatic values, inform the goals, contexts, and conditions

[489] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962; reprinted 2012

in which scientists (here, learning scientists) do their work, which shapes the directions of the tools they build. These research findings, instantiated in particular models, theories, and tools, make their way into practice—including by researchers embedded in practice (i.e., the self-described "learning engineers"[490]). Eventually, *successful* learning science projects implement their technology at scale, which then shapes the behaviors of other participants in this ecosystem – schools, teachers, and peer researchers. Those success stories are then fed back into what becomes valued as legitimate science—and what becomes seen as necessary requirements for both government and philanthropic funding in the next cycle. The result is a feedback loop that is constantly informing, shaping, and reinforcing a set of values and methods among learning scientists.

One can see this resistance to systemic change play out in the field of machine learning and its responses to critiques of ethics and fairness. For nearly six decades, humanists and critical theorists have levied critiques against techno-solutionist approaches to artificial intelligence[491,492]. Along the way, the field of machine learning labored on much as it always had, with funding calls and conference reviews continuing to ignore the harmful impacts of algorithmic systems. Following a confluence of high-profile revelations around discriminatory outcomes of machine learning systems[493,494,495,496] coinciding with these systems' widespread adoption in consumer products—and perhaps, broader societal changes—the conversation around ethics finally began to shift in the mid-2010s. But it wasn't until March 2019 that the National Science Foundation put out their first solicitation for grant proposals[497]. However, this call for grants falls squarely within the current techno-solutionist paradigm of thinking about fairness—making the assumption that such systems themselves may be fundamentally beneficial, and we must simply, as the NSF call puts it, "ensure benefits are broadly available across all segments of society." In their call, they specifically call for grants that develop novel technical methods for "detecting bias in systems" and "ensur[ing] fairness," without calling for research into interrogating or resisting the broader societal inequities that are instantiated and reproduced by algorithmic systems.

Kuhn's critique of scientific paradigms describes how hard it is for a scientific field to get out of this loop[498]. Disciplines are built on particular ways of knowing—think of bell hooks and how our ways of knowing are forged through inequitable relations of power. As a result, this normalizes what becomes defined as *science* at all. Conversely, these definitions also shape what becomes known as *not* science. As a result, researchers are expected to produce results within the normalized paradigm, doing the same work in the same

[490] Bror Saxberg. "Learning engineering: the art of applying learning science at scale". In: *Proceedings of the Fourth (2017) ACM Conference on Learning Scale*. ACM. 2017, pp. 1–1

[491] Norbert Wiener. *The human use of human beings: Cybernetics and society*. 320. Da Capo Press, 1988

[492] Terry Winograd, Fernando Flores, and Fernando F Flores. *Understanding computers and cognition: A new foundation for design*. Intellect Books, 1986

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

[494] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016

[495] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Proceedings of FAccT*. 2018, pp. 77–91

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

[497] National Science Foundation. *Program on Fairness in Artificial Intelligence in Collaboration with Amazon*. Accessed 2020-07-25. URL: https://bit.ly/2BvaTXo

[498] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962; reprinted 2012

structure of scientific practice. Unfortunately, anyone that tries to move outside of those boundaries, to produce radically different norms of knowledge, is immediately questioned, marginalized, and excluded from funding.

The self-reinforcing cycle is not produced by the people impacted by these systems, but by the powerful. Classroom teachers, for instance, even *if* they are given options about the types of educational technology to use in their classrooms, lack access to the levers of influence necessary to make fundamental, meaningful changes to the equity of learning science research. Indeed, teachers are often *most* subject to accountability for existing standards regardless of their much lower authority over institutional decision-making[499]. As Amy Ogan reports[500], while so-called "classroom sensing" technologies may ostensibly be designed to support teachers' self-reflective professional development, teachers have a "deep fear" that those tools will instead be co-opted by school administrators eager for awareness and control over teachers' behaviors. For graduate student researchers, the same applies. While they may be interested in changing fundamental power structures of educational technology research, to do so would put them at substantial professional risk. Taking on such risk is, itself, a privilege of those with safety nets and systemic power roles to draw on.

[499] Adam Kirk Edgerton and Laura M Desimone. "Mind the gaps: Differences in how teachers, principals, and districts experience college-and career-readiness policies". In: *American Journal of Education* 125.4 (2019), pp. 593–619

[500] Amy Ogan. "Reframing classroom sensing: promise and peril". In: *interactions* 26.6 (2019), pp. 26–32

Funding in Machine Learning for Education

Therefore, I focus my analysis on sources of funding that shape the nature of the research conducted, and highlight the choices that researchers may have in shaping these calls. In particular, I consider government agencies, corporations, and philanthropies.

Research funding is accessed primarily through the National Science Foundation (NSF) and the Department of Education, particularly the Institute for Education Sciences (IES). Smaller players in the learning sciences in particular include the National Institutes of Health (NIH) and DARPA, which is funded through the Department of Defense. While these organizations are much larger sources of grant funding for the sciences in general, they play a comparatively smaller role in shaping the landscape of education. The US Department of Education's Institute of Education Sciences, for instance, separates its grants into a series of Goals with incremental increases in the size of funding available. Early, exploratory proposals receive an order of magnitude less funding, capped at $700,000, but ostensibly do not require a track record of success to receive awards. Full-scale efficacy studies, on the other hand, can receive grants of up to $5,000,000 at a time, collaborations across multiple institutions,

numerous researchers and stakeholders, and many years of work.

While the federal funding pipeline is slow and relies on measurable evidence of each incremental change, the *corporate* world of educational technology software has historically needed little evidence at all. This results in extremely high variance between products: some have strong findings from research; others, by the time they make it to schools, bear little resemblance to the original experiments. Many products don't come from a research base to begin with, but are evaluated for success based on idiosyncratic measures, by district administration with informal, *ad hoc* training in data analytics[501].

[501] Alex J Bowers et al. "Education Leadership Data Analytics (ELDA): A White Paper Report". In: *ELDA Summit* (2019)

Another major source of funding for education research comes from the private sector. In capitalist economies, extremely high-net-worth captains of industry often use their wealth to fund philanthropic ventures. In the early twentieth century these projects included the funding of the arts by Rockefeller and the funding of public libraries by Carnegie. In modern times, this role is often played by organizations funded from the technology industry. Most visible among these organizations are the Bill & Melinda Gates Foundation and more recently, the Chan Zuckerberg Initiative.

While these organizations have sprawling operations that impact many different industries, because of all the factors listed above, they have a uniquely powerful role in defining the path of learning sciences research. In a generous reading, this role serves as a compromise between federal government funding and corporate investment in the education sector. Like government funders, philanthropies demand measurable results from their investments. Today there is an emphasis from essentially every philanthropic funder to combine a focus on measurable outcomes with an unusually tight timetable for experimentation, and in particular there is a need for grantees to demonstrate measurable efficacy results rapidly, often within a year or two of receiving funding. These "short-term wins" are often paired with longer-term strategic objectives of multi-year research agendas.

But unlike government research, these philanthropies fundamentally seek to turn research results into scalable, self-funded businesses that will turn into the corporate entities that sell their results to schools. This is a unique twist on the role of grant agencies in support of education. Corporate products as the primary driver of dissemination of research takes a fundamentally neoliberal view of how learning science and education reform can have an impact, focused on private sector investment[502,503,504]. This worldview has been described as *"philanthro-capitalism"* and defined as:

> *"the openness of personally profiting from charitable initiatives, an openness that deliberately collapses the distinction between public and private interests in order to justify increasingly concentrated levels of private gain."*[505]

[502] Joanne Barkan. "Plutocrats at work: How big philanthropy undermines democracy". In: *social research* 80.2 (2013), pp. 635–652

[503] Linsey McGoey. "Philanthrocapitalism and its critics". In: *Poetics* 40.2 (2012), pp. 185–199

[504] Ben Williamson. *Code Acts in Education: Re-Engineering Education*. 2020. URL: `https://nepc.colorado.edu/blog/re-engineering-education`

[505] Linsey McGoey. "Philanthrocapitalism and its critics". In: *Poetics* 40.2 (2012), pp. 185–199

This attitude is evident at every level of philanthropy in the learning sciences. The Chan Zuckerberg Initiative, for instance, is not a nonprofit organization but an LLC, inventing a new category of "for-profit philanthropy"[506]. This has blurred lines between the charitable and corporate roles of funders in education. In journalism, research has identified the lines of influence that philanthropic funders have over the journalistic organizations they fund[507]. In the learning sciences, often the same organizations that fund research into learning sciences (such as Chan Zuckerberg Initiative) also fund the conferences, journalism outlets (e.g., EdSurge), and publications that present the most optimistic and friendliest narrative around press releases and news of learning sciences innovations and product launches. This complex, deeply embedded role in the broader ecosystem gives philanthropies a remarkable influence over educational technology, having *"unprecedented power to shape the direction of research and development in education, by selecting and investing in programs that fit their personal vision."*[508]

[506] Ben Williamson. *Code Acts in Education: Re-Engineering Education*. 2020. URL: https://nepc.colorado.edu/blog/re-engineering-education

[507] Patrick Ferrucci and Jacob L Nelson. "The new advertisers: How foundation funding impacts journalism". In: *Media and Communication* 7.4 (2019), pp. 45–55

In fact, this undue influence of philanthro-capitalists in funding learning science research can be read as part of a larger tradition of disinvestment in public education in favor of the privatization of education (following Klein, and returning to the very beginning of this dissertation[509]). This trend further entrenches educational disparities along socioeconomic class lines.

[508] Ben Williamson. *Code Acts in Education: Re-Engineering Education*. 2020. URL: https://nepc.colorado.edu/blog/re-engineering-education

[509] Naomi Klein. *The shock doctrine: The rise of disaster capitalism.* Penguin Books, 2007

The Limits of Representation

One common response to these concerns is to acknowledge their validity, then shift to argue for their resolution through more representation and community voices in the decision-making process. This push for representation as a solution is a common and important step toward addressing issues of access and opportunity to the knowledge, practices, and spaces that are shaping algorithmic decision-making. However, in order to address concerns of socio-technical systems of oppression and inequities, this is not enough. Politics of representation makes strong assumptions about the idea that identity or representation shifts the epistemology and logic of socio-technical systems. In fact, a focus on identity or resources is not enough to address the non-material processes of power and oppression[510]. Unfortunately, these politics of representation do not address the normative disciplinary practices of producing knowledge and the hegemonic shaping of educational systems. Data science, and the training of it, is done within Western epistemological paradigms. Why should we expect that shifting representations will shift this hegemony?

[510] Iris Marion Young. *Justice and the Politics of Difference*. Princeton University Press, 1990

Part of the limitations of this inclusive design philosophy lies in

a critique of the focus on politics of representation. On one hand, there is a push to train more data scientists that are coming from underrepresented backgrounds; the theory of change here is that if those voices were part of the data science process, their perspectives would be included in the design of these algorithms and platforms. This also gets deployed in the way that communities and organizations are engaged: maybe research needs community members in the design process as a voice, as an attempt to shift representation, to make systems more equitable and less biased. One can see similar arguments in public policy[511] and community-engaged health research[512], as well as in the tradition of participatory design and community-driven co-design in human-computer interaction [513,514].

Parisi and Dixon-Román argue that this approach is based on a politics of inclusion[515]. Adding diversity to who is included at the table of decision making is an important initiative, but as Benjamin argues, "so much of what is routine, reasonable, intuitive, and codified reproduces unjust social arrangements, without ever burning a cross to shine light on the problem"[516]. These politics of inclusion fundamentally do not transform the norms and logic of reason that make up the epistemology of the system. Thus, a politics of inclusion in fact maintains and reifies a logic inherited by science and technology from a deeper, colonialist past. In other words, the issue at hand is not just about representation in voice and designers of existing systems, it is about recognizing the heritage of where those systems came from. Thus, adding (e.g., diverse team members) does not shift or change the underlying ways of thinking and knowing.

After all, who qualifies as an expert or a qualified voice at the table? The necessity for short-term outcomes in a particular format of efficacy study limits the researchers that are credibly able to apply for such funding: in particular, qualified applicants must have pre-existing research programs, participants for studies, and software ready to put into classrooms in a matter of months. For substantial later-stage support, that software must have been tested for a fairly narrow definition of successful learning gain. This introduces a hidden list of requirements into the top tier of funding: pre-existing relationships with a complex web of scientists, philanthropists, and educational settings, and a pre-existing agreement that those learning gains, measured typically on summative assessments of testable knowledge, are the appropriate metric of success. In today's philanthropic funding landscape for educational technology, a researcher cannot receive grants unless they are already enmeshed within the loop, having received grants before and built up social capital to receive further support while measuring what those with power have agreed to measure, using methods they have approved.

[511] Eric Corbett and Christopher A Le Dantec. "The problem of community engagement: Disentangling the practices of municipal government". In: *Proceedings of CHI*. 2018, pp. 1–13

[512] Joyce E Balls-Berry and Edna Acosta-Perez. "The use of community engaged research principles to improve health: community academic partnerships for research". In: *Puerto Rico health sciences journal* 36.2 (2017), p. 84

[513] Christina Harrington, Sheena Erete, and Anne Marie Piper. "Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–25

[514] Michael J Muller. "Participatory design: the third space in HCI". in: *The human-computer interaction handbook*. CRC press, 2007, pp. 1087–1108

[515] Ezekiel Dixon-Román and Luciana Parisi. "Data Capitalism, Sociogenic Prediction and Recursive Indeterminacies". In: *Public Plurality in an Era of Data Determinacy: Data Publics*. 2020

[516] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, 2019

To make the point more explicit: the requirements for funding and, therefore, ability to execute on a vision are structured for exclusion. They reinforce existing hierarchies in learning science and technology. The researchers that are qualified for new grants—at all but the earliest exploratory stages—are the researchers that were previously funded, and the projects with evidence that fits the funding agencies' targets are the projects they themselves have previously funded. Inclusion of a few select voices *into* this process is insufficient to change this feedback loop. Funding agencies pride themselves on their goals of funding innovation, and on finding research that is cutting-edge. But the template for innovation has been public and highly visible for years now. Funders do alter their criteria nominally from year to year, but the underlying premises have not changed. For at least two decades, there has been a focus on measurable outcomes and efficacy on standardized, summative assessments. Funders have made clear that the promising, scalable modalities for these approaches—those worthy of funding—are technological solutions, developed by companies that match the goals of philanthrocapitalists. If these solutions are where we place our optimism for deep-seated social issues, then we are asserting that those same organizations that led us to this point ought to choose the way to unwind their own historical impact on curriculum and learning.

Alternate Futures

So let's end with positive possibilities. What choices are available for active, inclusive, equitable change to how we do our work? All of these scholars I've cited have given us tools for explanation even deeper and broader than the social framework I've advanced. Only through engaging with these frameworks am I able to acknowledge the influence of external goals on the broader direction of the communities, products, and technologies I study. Do they tell us anything about how might we re-imagine a *better* future, one where defensible explanations are made on behalf of algorithms and tools that do not buy into these inequitable structures?

Policies and practices present particular visions of the world. Technologies then instantiate these visions and values into algorithmic systems that further entrench them. Theorists and critics have called to remake the financial world of machine learning research in a more just and more equitable way. I will not suddenly, at the end of this dissertation, be offering solutions to the core challenges of systemic oppression. But I will offer two potential ways forward: one that would entail changes to existing systems, and another that calls for transformative change.

New Research Priorities

First, I propose a change in current funding models, to better prioritize research that investigates how to make systemic social change, rather than relying on techno-solutionist thinking or on narrow experimental studies reporting incremental results. This is a natural extension of my more narrow call for social, non-causal account of explanation over a technical, introspective, causal account. This can be done through funding to researchers from marginalized groups. But allocation of funding to those individuals cannot work in isolation; allocation of resources within the existing feedback loop cannot be the extent of funding agencies' efforts towards addressing equity. It must be accompanied by changed expectations for the types of research that is publishable, towards research that is more participatory and more community-based.

This would mean requiring that research teams involve members of marginalized stakeholder groups as meaningful, co-equal members of the research team, or as advisors to the research projects, whose voices are given equal weight. That is, these stakeholders should be involved in framing the goals right from the start. When appropriate, they should be able to say that the research should not continue, if there is no way to make it equitable. This is in contrast to simply offering tokenized feedback to an already finished project, or a project on a fixed trajectory that can only be altered at the margins[517]. Addressing inequity through participatory design will require reconsideration of what it means for stakeholders to participate in research. Such work will also need to avoid overburdening marginalized communities, by compensating them adequately for the full scope of their newly expanded participation.

The field of machine learning is grappling with the complex contradictions in this path forward. Despite over 90 values statements[518] for ethical machine learning produced by large technology companies and government agencies , machine learning systems continue to be developed that discriminate and perpetuate injustice. While there may be legitimate organizational reasons why technology companies have been unable to put these principles into practice[519,520,521,522], others have been skeptical of intentions.

[517] Sherry R Arnstein. "A ladder of citizen participation". In: *Journal of the American Institute of planners* 35.4 (1969), pp. 216–224

[518] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1 (2019), pp. 389–399

[519] Kenneth Holstein et al. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of CHI*. 2018

[520] Michael Veale, Max Van Kleek, and Reuben Binns. "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making". In: *Proceedings of CHI*. 2018, pp. 1–14

[521] Luke Stark and Anna Lauren Hoffmann. "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture". In: *Journal of Cultural Analytics* (May 2019)

[522] Michael A Madaio et al. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI". in: *Proceedings of CHI*. 2020, pp. 1–14

Some critics describe these ethical statements as "fairwashing" or "ethics-washing," developing principles and value statements for ethical machine learning while doing little to change the nature of such technology, or the organizational processes that led to their design[523]. Mona Sloane has described this as a "smokescreen" for business as usual[524]. Thus, while it may be worthwhile to begin with changing the priorities and funding calls from the usual suspects of funding agencies to prioritize marginalized community interests and other ways of knowing, it will not be enough to transform fundamentally unjust systems.

One direction to look is what Mariam Asad et al. have referred to as "academic accomplices", or scholars whose research is designed to support the already ongoing justice work on communities [525]. Truly empowered participation by marginalized groups in these research agendas requires the ability to change ways of knowing what "evidence" looks like in learning science research. It will require a willingness to change the definition of efficacy and success in an intervention, based on what communities want and need. It will require looking at different disciplines, outside of the techno-solutionist mindset, and instead towards the voices of community organizers, social justice scholars, critical theorists and others. This will require a humility on the part of researchers and funders, a willingness to be part of a change that may leave them with less of a grip on power at the end of the funding process than they had at the start.

What we may learn from such engagement is that for some tasks, the current approach is inherently unjust and in these cases, the research should simply cease – "the implication is *not* to design"[526]. Os Keyes describes this situation in their critique of data science as rigid quantification that is intrinsically at odds with safety for those at the margins[527]. In such cases, injustices are inextricably embedded in the proposed technical solutions. For these types of tasks or methods, no amount of good explanation will uproot the underlying inequity.

[523] Elettra Bietti. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy". In: *Proceedings of FAccT*. 2020, pp. 210–219

[524] Mona Sloane. "Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice". In: *Weizenbaum Conference 2019 "Challenges of Digital Inequality — Digital Education, Digital Work, Digital Life"*. 2019

[525] Mariam Asad et al. "Academic Accomplices: Practical Strategies for Research Justice". In: *Proceedings of DIS*. 2019, pp. 353–356

[526] Eric PS Baumer and M Six Silberman. "When the implication is not to design (technology)". In: *Proceedings of CHI*. 2011, pp. 2271–2274

[527] Os Keyes. "Counting the Countless: Why data science is a profound threat for queer people". In: *Real Life* 2 (2019). URL: https://reallifemag.com/counting-the-countless/

Justice in Algorithmic Decision-Making Research

My final argument would be for machine learning researchers to adopt a *design justice* approach to algorithmic decision-making research. As proposed by Sasha Costanza-Chock, design justice is a *"framework for analysis of how design distributes benefits and burdens between various groups of people"*[528]. This involves interrogating the values encoded into designed systems, meaningfully involving members of marginalized and impacted communities in the design of systems, and questioning the narratives, sites, and pedagogies around design.

[528] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. MIT Press, 2020

To do this well, though, will require the communities I am a part of, in NLP, machine learning, and education technology to do some work. It requires us to learn, adopt, and promote theories and methods that critique the very power structures that have enabled our own research. Thus, a turn towards design justice may need to start by researchers, not funders. Changing a research community means changing practices, but also their public engagement, their calls for participation, review criteria, and awards of the journals and conferences they support. One may look to the medical field as an example here, with recent calls for community-engaged research leveraging journal reviews, grant funding, and universities' tenure and promotion as mechanisms for promoting a change in research values and methods[529].

The most robust version of this would involve the field as a whole changing our incentives. These incentives include the kinds of grants and funding that learning scientists pursue and accept, the collaborators they choose, the role of stakeholders in their research, and their engagement with the impact and legacy of their research. This engagement would continue after its dissemination and deployment in school systems, and necessarily prioritize a longer-term view over the current, more incremental status quo.

As a field, our vision for the kinds of technologies that can be built are shaped by what Sang-Hyun Kim and Sheila Jasanoff have called "*socio-technical imaginaries*:"

> *collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology*[530].

The socio-technical imaginaries of our current field are shaped and circumscribed by the history of technology as it is today – but they can be remade. To do this, we are inspired by Keyes et al.'s call for counterpower, or what they call emancipatory autonomy in human-computer interaction, or, more simply, anarchist HCI[531]. This involves fostering community-appropriate and community-determined research and design, both between researchers and stakeholders as well as within the technical research community itself. This move for counterpower would involve giving everyone, not simply a privileged few, the means to shape the forms of socio-technical educational systems. There are echoes of this in the calls to "democratize" machine learning from corporations[532] and researchers[533], which mostly seems to mean providing open-source tools for developing machine learning models (e.g., TensorFlow[534]). However, in practice, while platforms like TensorFlow may be freely available, simply giving more people access to them will do little to change the fun-

[529] Joyce E Balls-Berry and Edna Acosta-Perez. "The use of community engaged research principles to improve health: community academic partnerships for research". In: *Puerto Rico health sciences journal* 36.2 (2017), p. 84

[530] Sheila Jasanoff and Sang-Hyun Kim. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press, 2015

[531] Os Keyes, Josephine Hoy, and Margaret Drouhard. "Human-Computer Insurrection: Notes on an Anarchist HCI". in: *Proceedings of CHI*. 2019, pp. 1–13

[532] Microsoft. *Democratizing AI - Stories*. `https://news.microsoft.com/features/democratizing-ai/`. (Accessed on 05/01/2020). 2016

[533] Erwan Moreau, Carl Vogel, and Marguerite Barry. "A paradigm for democratizing artificial intelligence research". In: *Innovations in Big Data Mining and Embedded Knowledge*. Springer, 2019, pp. 137–166

[534] Martın Abadi et al. "Tensorflow: A system for large-scale machine learning". In: *USENIX Symposium on Operating Systems Design and Implementation*. 2016, pp. 265–283

damentally inequitable paradigms in which they are used. Instead, a more radical version of the democratization of learning technologies might reflect Paolo Freire's vision of a *liberatory* pedagogy that allows learners to pose the problems that are essential for their lives, and learn in ways that are meaningful and effective for them[535]. There is a rich lineage of this resistance to centralized educational hierarchy in education, including the critical theorist Ivan Illich, who famously called for "de-schooling society"[536], or Eli Meyerhoff's call for new "modes of study" beyond education[537].

In artificial intelligence more broadly, acts of resistance or refusal are becoming ever more common. For instance, communities have organized against computer vision used in public housing projects[538], and the Stop LAPD Spying Coalition,[539] has successfully organized to ban predictive policing algorithms in Los Angeles. One can also look to individual acts of resistance to harmful algorithmic systems, such as masks designed to resist facial recognition[540] and tools to obfuscate advertising algorithms[541]. However, these are acts of resistance and refusal to systems that are already designed and deployed, and likely already causing harm in the world.

Research methods that fit cleanly into the existing body of literature have resulted in reproductions of existing, inequitable brick-and-mortar education systems. We have built algorithms that reflect back the inequities that are already in place. This is not sufficient for creating new socio-technical imaginaries, or more liberatory forms of algorithmic decision-makers. As a field, we might look to methods from feminist speculative design[542,543] and critical design[544,545], which endeavor to provoke and problematize[546], to envision possible futures, both positive and negative, and to have bold visions for bringing these futures about. This might involve using critical design methods to provoke and problematize foundational assumptions in machine learning, including what should be learned and what it means to have learned it – to challenge these legacy views of what makes up a "good" model and a good decision.

Conclusion

All of this may require rethinking the standard machine learning research and design lifecycle to prioritize equity and justice[547] and to involve (and empower) stakeholders from marginalized communities throughout this lifecycle. This may entail a radical transformation of the "institutionally stabilized"[548] structures and incentives of the current system. It may require teaching students about the history and legacies of educational injustice, critical theory, and broader societal systems of oppression. And, it may require machine learning

[535] Paulo Freire. *Pedagogy of the oppressed*. Bloomsbury publishing USA, 1970

[536] Ivan Illich. *Deschooling society*. Penguin Group Limited, 1973

[537] Eli Meyerhoff. *Beyond Education: Radical Studying for Another World*. U of Minnesota Press, 2019

[538] Michele Gilman. *Voices of the Poor Must Be Heard in the Data Privacy Debate - JURIST - Commentary - Legal News & Commentary*. `https://www.jurist.org/commentary/2019/05/voices-of-the-poor-must-be-heard-in-the-data-privacy-debate/`. (Accessed on 05/01/2020). 2019

[539] https://stoplapdspying.org/

[540] Elise Thomas. *How to hack your face to dodge the rise of facial recognition tech |WIREDUK*. `https://www.wired.co.uk/article/avoid-facial-recognition-software`. (Accessed on 05/01/2020). 2019

[541] https://adnauseam.io/

[542] Anthony Dunne and Fiona Raby. *Speculative everything: design, fiction, and social dreaming*. MIT press, 2013

[543] Luiza Prado de O Martins. "Privilege and oppression: Towards a feminist speculative design". In: *Proceedings of DRS* (2014), pp. 980–990

[544] Shaowen Bardzell et al. "Critical design and critical theory: the challenge of designing for provocation". In: *Proceedings of DIS*. 2012, pp. 288–297

[545] Jeffrey Bardzell and Shaowen Bardzell. "What is" critical" about critical design?" In: *Proceedings of CHI*. 2013, pp. 3297–3306

[546] Laura Forlano and Anijo Mathew. "From design fiction to design friction: Speculative and participatory design of values-embedded urban technology". In: *Journal of Urban Technology* 21.4 (2014), pp. 7–24

[547] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. MIT Press, 2020

[548] Sheila Jasanoff and Sang-Hyun Kim. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press, 2015

researchers to become accomplices[549] with members of marginalized communities, to avoid reproducing existing power dynamics in their participatory design methods and move towards fostering counter-power in machine learning systems deployed in and for schools. All of this would require a radical re-envisioning of a more just, equitable, and liberatory system for learning science.

These problems are complex and interwoven and resist straightforward analyses and answers. Thinking about how your work is situated within a socio-technical system of self-reinforcing feedback loops is exhausting. Acknowledging that system and getting back to work is harder still. There's just no time to question the epistemologies of existing work and grants, much less find the space to actively support an activist, radical transformation.

But research on the underlying technologies and decision-making processes in education technology is a hint at the future of education policy itself. As I have shown throughout this dissertation, natural language processing is at the heart of many of these technologies. Language technologies research shapes real schools. Right now, as the very notion of *public education* is in flux in America, we are on the precipice of rapid and fundamental change and reform that will place technology in a front and center role. We should expect a rapid acceleration of the speed that our research is dropped into the classroom (or in today's remote world, the Zoom room).

Teachers, principals, and parents will have needs. Education companies looking to scale rapidly will have goals. The students themselves will have questions! We have so much opportunity to do the right kind of work as we introduce our technology, its benefits, its drawbacks, and its requirements. And the crisis of the COVID-19 pandemic, if nothing else, brings out a potential to break norms.

Maybe this crisis can be a reset point. Coming from positions of affluence and prestige in the tech industry, we are the ones with the privilege to break the cycle of technologies that do not really communicate what they do, or whose power they reinforce. As a community of researchers, there's so much we can collectively do to transform the trajectory of the field. The barriers to participation in the field of natural language processing and education technology are created by and reinforced by us, the researchers. The definition of what it means to build a "good" algorithm" is made up by us. This year our society have been undergoing an enormous rethinking of what it means to be technology-centric and enabled in the classroom, how quickly we should adopt tools, and who we should look to for help. Those shifts have all put us in an even greater position of responsibility than we had before. I believe it's necessary that we look up from our models and look around at the social environment where our tools are being

[549] Mariam Asad et al. "Academic Accomplices: Practical Strategies for Research Justice". In: *Proceedings of DIS*. 2019, pp. 353–356

deployed. It's necessary that we take that opportunity to reconnect with the real-world context of our work, and intentionally choose the priorities of our research. This choice includes picking the conflicts and discourses where algorithmic decision-making will step in, reshaping the path of the human decision-making it seeks to model. When the model does step in, we'll have made our choice: whose data we've included, whose decision-making we're emulating, what kind of intervention we're letting our automated tools make, and how much we know about that intervention and the ripple effects it is likely to have. Other disciplines have grappled with the responsibility they shoulder with their technological advances; I hope we do the same.

List of Publications

The work presented in this thesis spans a variety of interdisciplinary fields and major components of each chapter have already been published in peer-reviewed venues; those that remain are either under review or ready for development into publishable work in the 2020-2021 academic year.

The Philosophy of Explanation is based largely on a collaboration with researchers at the University of Kentucky, and was first published in LREC 2020 as *"Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models"*, coauthored with Christopher Grimsley and Julia R.S. Bursten.

All sections of ***Wikipedia Deletion Debates*** draw from two publications from 2019, coauthored with my advisor, Alan Black. First, at the ACL Workshop on Computational Social Science, we published details of the classification model described in ***Learning to Predict Decisions***, as *"Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making"*. The broader literature review from ***Context and Background***, as well as the contents of *Exploring and Explaining Decisions*, were published at the ACM SIGCHI Conference on Computer-Supported Collaborative Work (CSCW) as *"Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model"*.

Most of my work on automated essay scoring was completed in spring and summer of 2020 and has not yet been published in peer reviewed venues, with one exception. ***Evaluating Neural Models*** was originally published at the ACL Workshop on Innovative Uses of NLP for Building Educational Applications, as *"Should You Fine-Tune BERT for Automated Essay Scoring?"* The contents of the next two chapters, ***Training and Auditing DAACS*** and ***Explaining Essay Structure***, have been prepared for submission to the journal *Assessing Writing*, as *"Five-Paragraph Essays and Fair Automated Scoring in Online Higher Education"*. I expect this review process to be conducted in fall 2020. The topic modeling analysis of ***Explaining Essay Content*** is suitable for publication in a computer science venue focused on human-computer interaction, such as CSCW, and will likely be prepared for submission in the 2020-2021 academic year. Each of these publications is coauthored with Heidi Andrade, Jason Bryer, and Angela Lui, the leaders of the DAACS project, and with Alan Black.

The final section, ***Takeaways***, draws on multiple prior publica-

tions. ***Defensible Explanations*** draws on my work with Diyi Yang, first published as *"Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities."* in the ACM CHI conference, coauthored with Robert Kraut, Tenbroeck Smith, and Dan Jurafsky. It also proposes future work in philosophy of science that will likely be part of Christopher Grimsley's doctoral dissertation, and which will result in future publications on which I hope to collaborate. ***Confronting Inequity*** draws on collaborations with multiple collaborators. First, in the 2019 BEA workshop at ACL, I published *"Equity Beyond Bias in Language Technologies for Education"*, a nascent form of the chapter, in collaboration with Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, and Ezekiel Dixon-Román. A subset of us then continued discussing the implications of that workshop paper, and in early 2020, Michael Madaio, Ezekiel Dixon-Román, and I partnered with new contributor Su Lin Blodgett to submit *"Confronting Inherent Inequities in AI for Education"* to the International Journal on AI for Education. That article is currently under revision for a second round of peer review in fall 2020, and I expect it to be published after revisions in 2021.

Bibliography

[1] Martın Abadi et al. “Tensorflow: A system for large-scale machine learning”. In: *USENIX Symposium on Operating Systems Design and Implementation*. 2016, pp. 265–283.

[2] Ashraf Abdul et al. “Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda”. In: *Proceedings of CHI*. 2018, pp. 1–18.

[3] Diana Akhmedjanova et al. “Validity and Reliability of the DAACS Writing Assessment”. In: *Proceedings of NCME*. 2019.

[4] Khalid Al Khatib et al. “Modeling Deliberative Argumentation Strategies on Wikipedia”. In: *Proceedings of ACL*. Vol. 1. 2018, pp. 2545–2555.

[5] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. “Automatic Text Scoring Using Neural Networks”. In: *Proceedings of ACL*. 2016, pp. 715–725.

[6] Dimitris Alikaniotis and Vipul Raheja. “The Unreasonable Effectiveness of Transformer Language Models in Grammatical Error Correction”. In: *Proceedings of BEA*. 2019, pp. 127–133.

[7] H Samy Alim and Geneva Smitherman. *Articulate while Black: Barack Obama, language, and race in the US*. Oxford University Press, 2012.

[8] Walter R Allen et al. “From Bakke to Fisher: African American Students in US Higher Education over Forty Years”. In: *RSF: The Russell Sage Foundation Journal of the Social Sciences* 4.6 (2018), pp. 41–72.

[9] Mike Ananny and Kate Crawford. “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability”. In: *New Media & Society* 20.3 (2018), pp. 973–989.

[10] Anne H Anderson et al. “The HCRC map task corpus”. In: *Language and speech* 34.4 (1991), pp. 351–366.

[11] Heidi L Andrade, Ying Du, and Xiaolei Wang. "Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing". In: *Educational Measurement: Issues and Practice* 27.2 (2008), pp. 3–13.

[12] DA Angeli, Sheryl Brahnam, and Peter Wallis. "Abuse: The darker side of human computer interaction". In: *Interact 2005*. 2005, pp. 91–92.

[13] Julia Angwin et al. "Machine bias". In: *ProPublica, May* 23 (2016).

[14] Arthur N Applebee and Judith A Langer. "The state of writing instruction in America's schools: What existing data tell us". In: *Albany, NY: Center on English Learning and Achievement* (2006).

[15] Pablo Aragón et al. "Generative models of online discussion threads: state of the art and research challenges". In: *Journal of Internet Services and Applications* 8.1 (2017), p. 15.

[16] Sherry R Arnstein. "A ladder of citizen participation". In: *Journal of the American Institute of planners* 35.4 (1969), pp. 216–224.

[17] Mariam Asad et al. "Academic Accomplices: Practical Strategies for Research Justice". In: *Proceedings of DIS*. 2019, pp. 353–356.

[18] Yigal Attali. "Validity and Reliability of Automated Essay Scoring". In: *Handbook of automated essay evaluation: Current applications and new directions* (2013), p. 181.

[19] Yigal Attali and Jill Burstein. "Automated Essay Scoring with e-Rater® V. 2.0". In: *ETS Research Report Series* 2 (2004).

[20] Ben Backes, Harry J Holzer, and Erin Dunlop Velez. "Is it worth it? Postsecondary education and labor market outcomes for the disadvantaged". In: *IZA Journal of Labor Policy* 4.1 (2015), p. 1.

[21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *Proceedings of ICLR*. 2015.

[22] Ryan Shaun Baker and Paul Salvador Inventado. "Educational data mining and learning analytics". In: *Learning analytics*. Springer, 2014, pp. 61–75.

[23] Joyce E Balls-Berry and Edna Acosta-Perez. "The use of community engaged research principles to improve health: community academic partnerships for research". In: *Puerto Rico health sciences journal* 36.2 (2017), p. 84.

[24] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. "Gender identity and lexical variation in social media". In: *Journal of Sociolinguistics* 18.2 (2014), pp. 135–160.

[25] David Bamman, Brendan O'Connor, and Noah A Smith. "Learning latent personas of film characters". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 352–361.

[26] World Bank. *World Bank Education and COVID-19*. `https://bit.ly/2yZwGFa`. Accessed 2020-08-01.

[27] Jeffrey Bardzell and Shaowen Bardzell. "What is" critical" about critical design?" In: *Proceedings of CHI*. 2013, pp. 3297–3306.

[28] Shaowen Bardzell et al. "Critical design and critical theory: the challenge of designing for provocation". In: *Proceedings of DIS*. 2012, pp. 288–297.

[29] Joanne Barkan. "Plutocrats at work: How big philanthropy undermines democracy". In: *social research* 80.2 (2013), pp. 635–652.

[30] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of ACL*. 2014, pp. 238–247.

[31] Joost Bastings, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables". In: *Proceedings of ACL*. 2019.

[32] Robert W Batterman. *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press, 2001.

[33] Robert W Batterman and Collin C Rice. "Minimal model explanations". In: *Philosophy of Science* 81.3 (2014), pp. 349–376.

[34] Eric PS Baumer and M Six Silberman. "When the implication is not to design (technology)". In: *Proceedings of CHI*. 2011, pp. 2271–2274.

[35] Sian Bayne. "Teacherbot: interventions in automated teaching". In: *Teaching in Higher Education* 20.4 (2015), pp. 455–467.

[36] Daniel J Beal et al. "Cohesion and performance in groups: A meta-analytic clarification of construct relations." In: *Journal of applied psychology* 88.6 (2003), p. 989.

[37] Julia B Bear and Anita Williams Woolley. "The role of gender in team collaboration and performance". In: *Interdisciplinary science reviews* 36.2 (2011), pp. 146–153.

[38] Nadia Behizadeh. "Realizing powerful writing pedagogy in US public schools". In: *Pedagogies: An International Journal* 14.4 (2019), pp. 261–279.

[39] bell hooks bell. *Teaching community: A pedagogy of hope*. Vol. 36. Psychology Press, 2003.

[40] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, 2019.

[41] Cynthia L Bennett and Os Keyes. "What is the Point of Fairness? Disability, AI and The Complexity of Justice". In: *Workshop on AI Fairness for People with Disabilities at ACM SIGACCESS Conference on Computers and Accessibility*. 2019.

[42] Elettra Bietti. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy". In: *Proceedings of FAccT*. 2020, pp. 210–219.

[43] Or Biran and Courtenay Cotton. "Explanation and justification in machine learning: A survey". In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. 2017, p. 1.

[44] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[45] Su Lin Blodgett et al. "Language (technology) is power: The need to be explicit about NLP harms". In: *Proceedings of the Annual Meeting of the Association for Computational Lingustics (ACL)*. 2020.

[46] Alisa Bokulich. "Distinguishing explanatory from non-explanatory fictions". In: *Philosophy of Science* 79.5 (2012), pp. 725–737.

[47] Alisa Bokulich. "How scientific models can explain". In: *Synthese* 180.1 (2011), pp. 33–45.

[48] Alisa Bokulich. "Searching for Noncausal Explanations in a Sea of Causes". In: *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations* (2018), p. 141.

[49] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Proceedings of NeurIPS*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 4349–4357.

[50] Erik Borg. "Local plagiarisms". In: *Assessment & Evaluation in Higher Education* 34.4 (2009), pp. 415–426.

[51] Pierre Bourdieu. *Homo academicus*. Stanford University Press, 1988.

[52] Alex J Bowers et al. "Education Leadership Data Analytics (ELDA): A White Paper Report". In: *ELDA Summit* (2019).

[53] Jennie E Brand and Yu Xie. "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education". In: *American sociological review* 75.2 (2010), pp. 273–302.

[54] Jason Bryer and Angela M. Lui. "Efficacy of the Diagnostic Assessment and Achievement of College Students on Multiple Success Indicators". In: *Proceedings of AERA* (2019).

[55] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Proceedings of FAccT*. 2018, pp. 77–91.

[56] Moira Burke and Robert Kraut. "Mopping up: modeling wikipedia promotion decisions". In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM. 2008, pp. 27–36.

[57] Jill Burstein, Martin Chodorow, and Claudia Leacock. "Automated essay evaluation: The Criterion online writing service". In: *AI Magazine* 25.3 (2004), p. 27.

[58] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. "Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia". In: *Proceedings of CHI*. ACM. 2008, pp. 1101–1110.

[59] Jessica McCrory Calarco. ""I need help!" Social class and children's help-seeking in elementary school". In: *American Sociological Review* 76.6 (2011), pp. 862–882.

[60] Martin Carnoy and Emma Garcia. "Five Key Trends in US Student Performance: Progress by Blacks and Hispanics, the Takeoff of Asians, the Stall of Non-English Speakers, the Persistence of Socioeconomic Gaps, and the Damaging Effect of Highly Segregated Schools." In: *Economic Policy Institute* (2017).

[61] Heather M Caruso and Anita Williams Woolley. "Harnessing the power of emergent interdependence to promote diverse team collaboration". In: *Diversity and groups*. Emerald Group Publishing Limited, 2008, pp. 245–266.

[62] Leyland Cecco. "Female Nobel prize winner deemed not important enough for Wikipedia entry". In: *The Guardian* (2018). Accessed 2020-08-01. URL: https://bit.ly/38YxvMt.

[63] Adrien Chen. "Cambridge Analytica and our lives inside the surveillance machine". In: *The New Yorker* 21 (2018), pp. 8–10.

[64] Jing Chen et al. "Building e-rater® Scoring Models Using Machine Learning Methods". In: *ETS Research Report Series* 2016.1 (2016), pp. 1–12.

[65] Gordon W Cheung and Rebecca S Lau. "Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models". In: *Organizational research methods* 11.2 (2008), pp. 296–325.

[66] Francisco Chiclana et al. "A statistical comparative study of different similarity measures of consensus in group decision making". In: *Information Sciences* 221 (2013), pp. 110–123.

[67] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. "Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse". In: *Proceedings of CHI*. 2020, pp. 1–13.

[68] Martin Chodorow and Jill Burstein. "Beyond essay length: evaluating e-rater®'s performance on toefl® essays". In: *ETS Research Report Series* 2004.1 (2004), pp. i–38.

[69] Alexandra Chouldechova and Aaron Roth. "The Frontiers of Fairness in Machine Learning". In: *Workshop on Fair Representations and Fair Interactive Learning at the Computing Community Consortium* (2018).

[70] Brian D Christens and Paula Tran Inzeo. "Widening the view: situating collective impact among frameworks for community-led change". In: *Community Development* 46.4 (2015), pp. 420–435.

[71] Jason Chuang, Christopher D Manning, and Jeffrey Heer. "Termite: Visualization techniques for assessing textual topic models". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 2012, pp. 74–77.

[72] Kevin Clark et al. "What Does BERT Look At? An Analysis of BERT's Attention". In: *Workshop on Blackbox NLP at ACL*. 2019.

[73] Kevin Cokley. "What do we know about the motivation of African American students? Challenging the" anti-intellectual" myth". In: *Harvard educational review* 73.4 (2003), pp. 524–558.

[74] Nicholas B Colvard, C Edward Watson, and Hyojin Park. "The Impact of Open Educational Resources on Various Student Success Metrics." In: *International Journal of Teaching and Learning in Higher Education* 30.2 (2018), pp. 262–276.

[75] William Condon. "Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?" In: *Assessing Writing* 18.1 (2013), pp. 100–108.

[76] Sam Corbett-Davies and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning". In: *Synthesis of tutorial presented at ICML.* (2018).

[77] Eric Corbett and Christopher A Le Dantec. "The problem of community engagement: Disentangling the practices of municipal government". In: *Proceedings of CHI*. 2018, pp. 1–13.

[78] Gonçalo M Correia, Vlad Niculae, and André FT Martins. "Adaptively Sparse Transformers". In: *Proceedings of EMNLP*. 2019, pp. 2174–2184.

[79] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. MIT Press, 2020.

[80] Elena Cotos. "Automated Writing Analysis for writing pedagogy: From healthy tension to tangible prospects". In: *Writing and Pedagogy* 6 (2015), p. 1.

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

[82] Kelley Cotter, Janghee Cho, and Emilee Rader. "Explaining the news feed algorithm: An analysis of the" News Feed FYI" blog". In: *Proceedings of CHI Extended Abstracts*. 2017.

[83] Tressie McMillan Cottom. *Lower ed: The troubling rise of for-profit colleges in the new economy*. New Press, The, 2017.

[84] Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. "Automated essay scoring with string kernels and word embeddings". In: *Proceedings of ACL*. 2018.

[85] Scott Crossley et al. "Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system". In: *Proceedings of AIED*. Springer. 2013.

[86] Anne Curzan. "Teaching the politics of standard English". In: *Journal of English Linguistics* 30.4 (2002), pp. 339–352.

[87] Zihang Dai et al. "Transformer-xl: Attentive language models beyond a fixed-length context". In: *Proceedings of ACL*. 2019.

[88] Linda Darling-Hammond and Channa M Cook-Harvey. "Educating the whole child: Improving school climate to support student success". In: *Palo Alto, CA: Learning Policy Institute* (2018).

[89] Henk W De Regt. "The epistemic value of understanding". In: *Philosophy of Science* 76.5 (2009), pp. 585–597.

[90] Simon DeDeo. "Group minds and the case of Wikipedia". In: *Human Computation* (2014).

[91] Sara Delamont, Odette Parry, and Paul Atkinson. "Critical mass and pedagogic continuity: studies in academic habitus". In: *British Journal of Sociology of Education* 18.4 (1997), pp. 533–549.

[92] Robin DeRosa and Scott Robison. "From OER to open pedagogy: Harnessing the power of open". In: *Open: The philosophy and practices that are revolutionizing education and science. London: Ubiquity Press.* 4.1 (2017), p. 0.

[93] Katie Derthick et al. "Collaborative sensemaking during admin permission granting in Wikipedia". In: *International Conference on Online Communities and Social Computing*. Springer. 2011, pp. 100–109.

[94] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of NAACL*. 2019.

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

[96] Adji B Dieng, Francisco JR Ruiz, and David M Blei. "Topic modeling in embedding spaces". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 439–453.

[97] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1 (2015), p. 114.

[98] Semire Dikli. "An overview of automated scoring of essays". In: *The Journal of Technology, Learning and Assessment* 5.1 (2006).

[99] Stephanie Dix. ""What did I change and why did I do it?" Young writers' revision practices". In: *Literacy* 40.1 (2006), pp. 3–10.

[100] Ezekiel Dixon-Román and Luciana Parisi. "Data Capitalism, Sociogenic Prediction and Recursive Indeterminacies". In: *Public Plurality in an Era of Data Determinacy: Data Publics*. 2020.

[101] Lucas Dixon et al. "Measuring and mitigating unintended bias in text classification". In: *Proceedings of AIES*. ACM. 2018, pp. 67–73.

[102] Fei Dong, Yue Zhang, and Jie Yang. "Attention-based recurrent convolutional neural network for automatic essay scoring". In: *Proceedings of CONLL*. 2017, pp. 153–162.

[103] Phil Dowe. "An empiricist defence of the causal account of explanation". In: *International Studies in the Philosophy of Science* 6.2 (1992), pp. 123–128.

[104] John P Dugan, Michelle L Kusel, and Dawn M Simounet. "Transgender college students: An exploratory study of perceptions, engagement, and educational outcomes". In: *Journal of College Student Development* 53.5 (2012), pp. 719–736.

[105] Anthony Dunne and Fiona Raby. *Speculative everything: design, fiction, and social dreaming*. MIT press, 2013.

[106] Adam Kirk Edgerton and Laura M Desimone. "Mind the gaps: Differences in how teachers, principals, and districts experience college-and career-readiness policies". In: *American Journal of Education* 125.4 (2019), pp. 593–619.

[107] Upol Ehsan et al. "Rationalization: A neural machine translation approach to generating natural language explanations". In: *Proceedings of AIES*. ACM. 2018, pp. 81–87.

[108] Peter Elbow. "Closing my eyes as I speak: An argument for ignoring audience". In: *College English* 49.1 (1987), pp. 50–69.

[109] Jannette Elwood. "Equity issues in performance assessment: The contribution of teacher-assessed coursework to gender-related differences in examination performance". In: *Educational Research and Evaluation* 5.4 (1999), pp. 321–344.

[110] William Emigh and Susan C Herring. "Collaborative authoring on the web: A genre analysis of online encyclopedias". In: *Proceedings of HICSS*. IEEE. 2005.

[111] Hans-Peter Erb, Antoine Bioy, and Denis J Hilton. "Choice preferences without inferences: Subconscious priming of risk attitudes". In: *Journal of Behavioral Decision Making* 15.3 (2002), pp. 251–262.

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

[113] Rong-En Fan et al. "LIBLINEAR: A library for large linear classification". In: *Journal of machine learning research* 9.Aug (2008), pp. 1871–1874.

[114] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. "Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input". In: *Proceedings of NAACL*. 2018, pp. 263–271.

[115] Dana R. Ferris. "Student Reactions to Teacher Response in Multiple-Draft Composition Classrooms". In: *TESOL Quarterly* 29.1 (1995), pp. 33–53.

[116] Patrick Ferrucci and Jacob L Nelson. "The new advertisers: How foundation funding impacts journalism". In: *Media and Communication* 7.4 (2019), pp. 45–55.

[117] James Fiacco, Elena Cotos, and Carolyn Rosé. "Towards Enabling Feedback on Rhetorical Structure with Neural Sequence Models". In: *Proceedings of LAK*. ACM. 2019, pp. 310–319.

[118] Peter W Foltz, Sara Gilliam, and Scott Kendall. "Supporting content-based feedback in on-line writing evaluation with LSA". In: *Interactive Learning Environments* 8.2 (2000), pp. 111–127.

[119] Kate Forbes-Riley and Diane Litman. "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor". In: *Speech Communication* 53.9-10 (2011), pp. 1115–1136.

[120] Denae Ford et al. "Paradise unplugged: Identifying barriers for female participation on stack overflow". In: *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM. 2016, pp. 846–857.

[121] Signithia Fordham and John U Ogbu. "Black students' school success: Coping with the "burden of 'acting white'"". In: *The urban review* 18.3 (1986), pp. 176–206.

[122] Laura Forlano and Anijo Mathew. "From design fiction to design friction: Speculative and participatory design of values-embedded urban technology". In: *Journal of Urban Technology* 21.4 (2014), pp. 7–24.

[123] Andrea Forte and Amy Bruckman. "Scaling consensus: Increasing decentralization in Wikipedia governance". In: *Proceedings of HICSS*. IEEE. 2008, pp. 157–157.

[124] Andrea Forte and Amy Bruckman. "Why do people write for Wikipedia? Incentives to contribute to open–content publishing". In: *Proceedings of Group* (2005), pp. 6–9.

[125] Andrea Forte, Vanesa Larco, and Amy Bruckman. "Decentralization in Wikipedia governance". In: *Journal of Management Information Systems* 26.1 (2009), pp. 49–72.

[126] Becky Francis et al. "University lecturers' perceptions of gender and undergraduate writing". In: *British Journal of Sociology of Education* 24.3 (2003), pp. 357–373.

[127] Paulo Freire. *Pedagogy of the oppressed*. Bloomsbury publishing USA, 1970.

[128] Michael Friedman. "Explanation and scientific understanding". In: *The Journal of Philosophy* 71.1 (1974), pp. 5–19.

[129] Dennis Friess and Christiane Eilders. "A systematic review of online deliberation research". In: *Policy & Internet* 7.3 (2015), pp. 319–339.

[130] Stephen Frosh, Ann Phoenix, and Rob Pattman. "The trouble with boys". In: *The Psychologist* 16.2 (2003), pp. 84–87.

[131] Andreas Fuster et al. "Predictably unequal? the effects of machine learning on credit markets". In: *The Effects of Machine Learning on Credit Markets (November 6, 2018)* (2018).

[132] Richard Stuart Geiger II. "Robots. txt: An Ethnographic Investigation of Automated Software Agents in User-Generated Content Platforms". PhD thesis. University of California, Berkeley, 2015.

[133] Dan Geiger, Thomas Verma, and Judea Pearl. "Identifying independence in Bayesian networks". In: *Networks* 20.5 (1990), pp. 507–534.

[134] R Stuart Geiger and Heather Ford. "Participation in Wikipedia's article deletion processes". In: *Proceedings of WikiSym*. 2011, pp. 201–202.

[135] Máirtın Mac an Ghaill. "'What about the boys?': schooling, class and crisis masculinity". In: *The Sociological Review* 44.3 (1996), pp. 381–397.

[136] Jim Giles. *Internet encyclopaedias go head to head*. 2005.

[137] Tarleton Gillespie. "The relevance of algorithms". In: *Media technologies: Essays on communication, materiality, and society* 167.2014 (2014), p. 167.

[138] Carol Gilligan and Ina Different Voice. "Psychological theory and womens development". In: *Cambridge, MA* (1993).

[139] Michele Gilman. *Voices of the Poor Must Be Heard in the Data Privacy Debate - JURIST - Commentary - Legal News & Commentary*. https://www.jurist.org/commentary/2019/05/voices-of-the-poor-must-be-heard-in-the-data-privacy-debate/. (Accessed on 05/01/2020). 2019.

[140] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. "Toward establishing trust in adaptive agents". In: *Proceedings of the International Conference on Intelligent User Interfaces*. 2008.

[141] Ella Glikson and Anita Williams Woolley. "Human trust in Artificial Intelligence: Review of empirical research". In: *Academy of Management Annals* ja (2020).

[142] Sara Goldrick-Rab. "Following their every move: An investigation of social-class differences in college pathways". In: *Sociology of Education* 79.1 (2006), pp. 67–79.

[143] Matthew C Gombolay et al. "Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams". In: *Autonomous Robots* 39.3 (2015).

[144] Hila Gonen and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". In: *Proceedings of NAACL*. 2019, pp. 609–614.

[145] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine* 38.3 (2017), pp. 50–57.

[146] Edmund W Gordon and Kavitha Rajagopalan. "Assessment for Teaching and Learning, Not Just Accountability". In: *The Testing and Learning Revolution*. Springer, 2016, pp. 9–34.

[147] Steve Graham and Karen Harris. "Writing Better: Effective Strategies for Teaching Students with Learning Difficulties." In: *Brookes Publishing Company* (2005).

[148] Douglas Grimes and Mark Warschauer. "Utility in a fallible tool: A multi-site case study of automated writing evaluation." In: *Journal of Technology, Learning, and Assessment* 8.6 (2010).

[149] Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models". In: *Proceedings of LREC*. 2020.

[150] Magnus Haake and Agneta Gulz. "Visual stereotypes and virtual pedagogical agents". In: *Journal of Educational Technology & Society* 11.4 (2008).

[151] J Richard Hackman. *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers, 2011.

[152] Aaron Halfaker. "Interpolating quality dynamics in wikipedia and demonstrating the keilana effect". In: *Proceedings of WikiSym*. ACM. 2017, p. 19.

[153] Aaron Halfaker et al. "The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline". In: *American Behavioral Scientist* 57.5 (2013).

[154] Andrew Hall, Loren Terveen, and Aaron Halfaker. "Bot Detection in Wikidata Using Behavioral and Other Informal Cues". In: *Proceedings of CSCW* (2018), p. 64.

[155] Longfei Han et al. "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes". In: *IEEE journal of biomedical and health informatics* 19.2 (2014), pp. 728–734.

[156] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. "Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions". In: *Proceedings of ACL*. 2020.

[157] Christina Harrington, Sheena Erete, and Anne Marie Piper. "Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–25.

[158] John Hayes and Linda Flower. "Identifying the Organization of Writing Processes". In: *Cognitive Processes in writing*. Ed. by L Gregg and E Teinber. Erlbaum, 1980.

[159] Bradi Heaberlin and Simon DeDeo. "The evolution of Wikipedia's norm network". In: *Future Internet* 8.2 (2016), p. 14.

[160] Neil T Heffernan and Cristina Lindquist Heffernan. "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching". In: *International Journal of Artificial Intelligence in Education* 24.4 (2014), pp. 470–497.

[161] Alex Helberg et al. "Teaching textual awareness with DocuScope: Using corpus-driven tools and reflection to support students' written decision-making". In: *Assessing Writing* 38 (2018), pp. 40–45.

[162] Tove Helldin et al. "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving". In: *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 2013, pp. 210–217.

[163] Carl G Hempel and Paul Oppenheim. "Studies in the Logic of Explanation". In: *Philosophy of science* 15.2 (1948), pp. 135–175.

[164] Monika Hengstler, Ellen Enkel, and Selina Duelli. "Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices". In: *Technological Forecasting and Social Change* 105 (2016).

[165] Derrick Higgins et al. "Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring". In: *arXiv preprint arXiv:1403.0801* (2014).

[166] Charles G Hill et al. "Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways?" In: *Proceedings of CHI*. ACM. 2017, pp. 6658–6671.

[167] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network". In: *stat* 1050 (2015), p. 9.

[168] Kevin Anthony Hoff and Masooda Bashir. "Trust in automation: Integrating empirical evidence on factors that influence trust". In: *Human factors* 57.3 (2015), pp. 407–434.

[169] Guy Hoffman and Cynthia Breazeal. "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 2007, pp. 1–8.

[170] Ronny Högberg. "Cheating as subversive and strategic resistance: vocational students' resistance and conformity towards academic subjects in a Swedish upper secondary school". In: *Ethnography and Education* 6.3 (2011), pp. 341–355.

[171] David Holbrook. "Native American ELL Students, Indian English, and the Title III Formula Grant". In: *Annual Bilingual/Multicultural Education Conference*. 2011.

[172] Kenneth Holstein et al. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of CHI*. 2018.

[173] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* (2017).

[174] Guangwei Hu and Jun Lei. "Chinese university students' perceptions of plagiarism". In: *Ethics & Behavior* 25.3 (2015), pp. 233–255.

[175] Yiqing Hua et al. "WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community". In: *Proceedings of EMNLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2818–2823. URL: http://aclweb.org/anthology/D18-1305.

[176] Anne H Charity Hudley and Christine Mallinson. *We Do Language: English Variation in the Secondary English Classroom*. Teachers College Press, 2013.

[177] Janet Shibley Hyde. "The gender similarities hypothesis." In: *American psychologist* 60.6 (2005), p. 581.

[178] Ivan Illich. *Deschooling society*. Penguin Group Limited, 1973.

[179] Jane Im et al. "Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments". In: *Proceedings of CSCW* (2018), p. 74.

[180] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of NAACL*. 2019.

[181] Sarthak Jain et al. "Learning to faithfully rationalize by construction". In: *Proceedings of ACL*. 2020.

[182] Sandra Jamieson. "Is it plagiarism or patchwriting? Toward a nuanced definition". In: *Handbook of academic integrity* (2016), pp. 503–518.

[183] Sheila Jasanoff and Sang-Hyun Kim. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press, 2015.

[184] Ali Javanmardi and Lu Xiao. "What's in the Content of Wikipedia's Article for Deletion Discussions?" In: *Proceedings of The Web Conference (WWW)*. 2019, pp. 1215–1223.

[185] Ganesh Jawahar, Benoıt Sagot, and Djamé Seddah. "What Does BERT Learn about the Structure of Language?" In: *Proceedings of ACL*. 2019, pp. 3651–3657.

[186] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. "Why differences make a difference: A field study of diversity, conflict and performance in workgroups". In: *Administrative science quarterly* 44.4 (1999), pp. 741–763.

[187] Rajiv S Jhangiani et al. "As Good or Better than Commercial Textbooks: Students' Perceptions and Outcomes from Using Open Digital and Open Print Textbooks." In: *Canadian Journal for the Scholarship of Teaching and Learning* 9.1 (2018), n1.

[188] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1 (2019), pp. 389–399.

[189] Adam C Johnson, Joshua Wilson, and Rod D Roscoe. "College student perceptions of writing errors, text quality, and author characteristics". In: *Assessing Writing* 34 (2017), pp. 72–87.

[190] Eric J Johnson et al. "Beyond nudges: Tools of a choice architecture". In: *Marketing Letters* 23.2 (2012), pp. 487–504.

[191] Elisabeth Joyce, Jacqueline C Pike, and Brian S Butler. "Rules and roles vs. consensus: Self-governed deliberative mass collaboration bureaucracies". In: *American Behavioral Scientist* 57.5 (2013), pp. 576–594.

[192] Dan Jurafsky and James H Martin. *Speech and language processing*. Vol. 3. Pearson London, 2014.

[193] Dongyeop Kang and Eduard Hovy. "xSLUE: A Benchmark and Analysis Platform for Cross-Style Language Understanding and Evaluation". In: *arXiv preprint arXiv:1911.03663* (2019).

[194] Harmanpreet Kaur et al. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning". In: *Proceedings of CHI*. 2020, pp. 1–14.

[195] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. "Learning The Difference That Makes A Difference With Counterfactually-Augmented Data". In: *Proceedings of ICLR*. 2020.

[196] Brian Keegan and Casey Fiesler. "The Evolution and Consequences of Peer Producing Wikipedia's Rules". In: *Proceedings of ICWSM* (2017).

[197] Brian Keegan and Darren Gergle. "Egalitarians at the gate: One-sided gatekeeping practices in social media". In: *Proceedings CSCW*. 2010, pp. 131–134.

[198] Ursula Kessels et al. "How gender differences in academic engagement relate to students' gender identity". In: *Educational Research* 56.2 (2014), pp. 220–229.

[199] Os Keyes. "Counting the Countless: Why data science is a profound threat for queer people". In: *Real Life* 2 (2019). URL: https://reallifemag.com/counting-the-countless/.

[200] Os Keyes. "The misgendering machines: Trans/HCI implications of automatic gender recognition". In: *Proceedings of CSCW* (2018).

[201] Os Keyes, Josephine Hoy, and Margaret Drouhard. "Human-Computer Insurrection: Notes on an Anarchist HCI". In: *Proceedings of CHI*. 2019, pp. 1–13.

[202] Kareem Khalifa. "The role of explanation in understanding". In: *The British Journal for the Philosophy of Science* 64.1 (2012), pp. 161–187.

[203] Philip Kitcher. "Explanatory unification". In: *Philosophy of science* 48.4 (1981), pp. 507–531.

[204] Aniket Kittur et al. "He says, she says: conflict and coordination in Wikipedia". In: *Proceedings of CHI*. ACM. 2007, pp. 453–462.

[205] Naomi Klein. "Screen New Deal". In: *The Intercept* (2020). Accessed 2020-08-01. `https://bit.ly/3dZJhXw`.

[206] Naomi Klein. *The shock doctrine: The rise of disaster capitalism.* Penguin Books, 2007.

[207] Pang Wei Koh and Percy Liang. "Understanding Black-box Predictions via Influence Functions". In: *International Conference on Machine Learning*. 2017, pp. 1885–1894.

[208] Olga Kovaleva et al. "Revealing the Dark Secrets of BERT". In: *Proceedings of EMNLP*. Vol. 1. 2019, pp. 2465–2475.

[209] Steve WJ Kozlowski and Daniel R Ilgen. "Enhancing the effectiveness of work groups and teams". In: *Psychological science in the public interest* 7.3 (2006), pp. 77–124.

[210] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8788–8790.

[211] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962; reprinted 2012.

[212] Carmen Lacave and Francisco J Díez. "A review of explanation methods for Bayesian networks". In: *The Knowledge Engineering Review* 17.2 (2002), pp. 107–127.

[213] Gloria Ladson-Billings. "From the achievement gap to the education debt: Understanding achievement in US schools". In: *Educational researcher* 35.7 (2006), pp. 3–12.

[214] Gloria Ladson-Billings. "Toward a theory of culturally relevant pedagogy". In: *American educational research journal* 32.3 (1995), pp. 465–491.

[215] Himabindu Lakkaraju et al. "Interpretable & explorable approximations of black box models". In: *Proceedings of KDD Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2017).

[216] Shyong K Lam, Jawed Karim, and John Riedl. "The effects of group composition on decision quality in a social production community". In: *Proceedings of Group*. ACM. 2010, pp. 55–64.

[217] Shyong K Lam et al. "WP: clubhouse?: an exploration of Wikipedia's gender imbalance". In: *Proceedings of WikiSym*. ACM. 2011, pp. 1–10.

[218] Marc Lange. "What makes a scientific explanation distinctively mathematical?" In: *The British Journal for the Philosophy of Science* 64.3 (2013), pp. 485–511.

[219] Min Kyung Lee et al. "Working with machines: The impact of algorithmic and data-driven management on human workers". In: *Proceedings of CHI*. 2015, pp. 1603–1612.

[220] Krittaya Leelawong and Gautam Biswas. "Designing learning by teaching agents: The Betty's Brain system". In: *International Journal of Artificial Intelligence in Education* 18.3 (2008), pp. 181–208.

[221] David Lehr and Paul Ohm. "Playing with the Data: What Legal Scholars Should Learn About Machine Learning". In: *UCDL Rev.* 51 (2017), p. 653.

[222] Tao Lei, Regina Barzilay, and Tommi Jaakkola. "Rationalizing Neural Predictions". In: *Proceedings of EMNLP*. 2016, pp. 107–117.

[223] John M Levine, Lauren B Resnick, and E Tory Higgins. "Social foundations of cognition". In: *Annual review of psychology* 44.1 (1993), pp. 585–612.

[224] Jiwei Li et al. "Visualizing and Understanding Neural Models in NLP". In: *Proceedings of NAACL*. 2016, pp. 681–691.

[225] Andrew Lih. "Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource". In: *Proceedings of the International Symposium on Online Journalism, 2004*. 2004.

[226] Brian Y Lim and Anind K Dey. "Assessing demand for intelligibility in context-aware applications". In: *Proceedings of the International Conference on Ubiquitous Computing*. ACM. 2009, pp. 195–204.

[227] Zachary C Lipton. "The mythos of model interpretability". In: *ICML Workshop on Human Interpretabiltiiy in Machine Learning* (2016).

[228] Frederick Liu and Besim Avci. "Incorporating Priors with Feature Attribution on Text Classification". In: *Proceedings of ACL*. 2019.

[229] Hui Liu, Qingyu Yin, and William Yang Wang. "Towards Explainable NLP: A Generative Explanation Framework for Text Classification". In: *Proceedings of NAACL*. 2019.

[230] Shusen Liu et al. "Visual interrogation of attention-based models for natural language inference and machine comprehension". In: *Proceedings of EMNLP*. 2018.

[231] Audre Lorde. "The master's tools will never dismantle the master's house". In: *Sister outsider: Essays and speeches* 1 (1984), pp. 10–14.

[232] Yin Lou, Rich Caruana, and Johannes Gehrke. "Intelligible models for classification and regression". In: *Proceedings of the ACM SIGKDD*. ACM. 2012, pp. 150–158.

[233] Anastassia Loukina, Nitin Madnani, and Klaus Zechner. "The many dimensions of algorithmic fairness in educational applications". In: *Proceedings of BEA*. 2019, pp. 1–10.

[234] Anastassia Loukina et al. "Using exemplar responses for training and evaluating automated speech scoring systems". In: *Proceedings of BEA*. 2018, pp. 1–12.

[235] Michael A Madaio et al. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI". In: *Proceedings of CHI*. 2020, pp. 1–14.

[236] Michael Madaio et al. "Confronting Inherent Inequities in AI for Education". In: *International Journal of Artificial Intelligence in Education* (Under review).

[237] Nitin Madnani et al. "Building better open-source tools to support fairness in automated scoring". In: *Proceedings of the Workshop on Ethics in Natural Language Processing at ACL*. 2017, pp. 41–52.

[238] Fiona Mao, Robert Mercer, and Lu Xiao. "Extracting imperatives from wikipedia article for deletion discussions". In: *Proceedings of the Workshop on Argumentation Mining at ACL*. 2014, pp. 106–107.

[239] John Markoff. "Essay-Grading Software Offers Professors a Break". In: *The New York Times* (Apr. 2013). (Accessed on 06-30-2020.) URL: https://nyti.ms/2BoUaoF.

[240] Ramon Antonio Martinez. *Spanglish is spoken here: Making sense ofSpanish-English code-switching and language ideologies in a sixth-grade English language arts classroom*. University of California, Los Angeles, 2009.

[241] Luiza Prado de O Martins. "Privilege and oppression: Towards a feminist speculative design". In: *Proceedings of DRS* (2014), pp. 980–990.

[242] Paul Kei Matsuda and Christine M Tardy. "Voice in academic writing: The rhetorical construction of author identity in blind manuscript review". In: *English for Specific Purposes* 26.2 (2007), pp. 235–249.

[243] Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. "Recognizing rare social phenomena in conversation: Empowerment detection in support group chatrooms". In: *Proceedings of ACL*. Vol. 1. 2013, pp. 104–113.

[244] Elijah Mayfield and Alan W Black. "Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–26.

[245] Elijah Mayfield and Alan W Black. "Should you Fine-Tune BERT for Automated Essay Scoring?" In: *Proceedings of BEA*. 2020.

[246] Elijah Mayfield and Alan W Black. "Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making". In: *Workshop on Natural Language Processing + Computational Social Science at NAACL*. 2019.

[247] Elijah Mayfield et al. "Computational representation of discourse practices across populations in task-based dialogue". In: *Proceedings of the International Conference on Intercultural Collaboration*. ACM. 2012, pp. 67–76.

[248] Elijah Mayfield et al. "Five-Paragraph Essays and Fair Automated Scoring in Online Higher Education". In: *Assessing Writing*. Under review.

[249] Iain McCowan et al. "The AMI meeting corpus". In: *Proceedings of the Conference on Methods and Techniques in Behavioral Research*. Vol. 88. 2005, p. 100.

[250] Linsey McGoey. "Philanthrocapitalism and its critics". In: *Poetics* 40.2 (2012), pp. 185–199.

[251] Joseph Edward McGrath. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[252] Danielle S McNamara et al. "A hierarchical classification approach to automated essay scoring". In: *Assessing Writing* 23 (2015), pp. 35–59.

[253] Amanda Menking and Ingrid Erickson. "The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia". In: *Proceedings of CHI*. ACM. 2015, pp. 207–210.

[254] Mostafa Mesgari et al. ""The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia". In: *Journal of the Association for Information Science and Technology* 66.2 (2015), pp. 219–245.

[255] Robinson Meyer. "Everything We Know About Facebook's Secret Mood Manipulation Experiment". In: *The Atlantic* (2014). Accessed 2020-08-01. https://bit.ly/30l57kA.

[256] Eli Meyerhoff. *Beyond Education: Radical Studying for Another World*. U of Minnesota Press, 2019.

[257] Microsoft. *Democratizing AI - Stories*. https://news.microsoft.com/features/democratizing-ai/. (Accessed on 05/01/2020). 2016.

[258] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* (2018).

[259] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences". In: *Proceedings of ICJAI Workshop on Explainable AI*. 2017.

[260] Tristan Miller. "Essay assessment with latent semantic analysis". In: *Journal of Educational Computing Research* 29.4 (2003), pp. 495–512.

[261] Frances J Milliken, Caroline A Bartel, and Terri R Kurtzberg. "Diversity and creativity in work groups". In: *Group creativity: Innovation through collaboration* (2003), pp. 32–62.

[262] Nicole Mirnig et al. "To err is robot: How humans assess and act toward an erroneous social robot". In: *Frontiers in Robotics and AI* 4 (2017), p. 21.

[263] Brent Daniel Mittelstadt et al. "The ethics of algorithms: Mapping the debate". In: *Big Data & Society* 3.2 (2016).

[264] M Möhlmann and L Zalmanson. "Hands on the wheel: Navigating algorithmic management and Uber drivers'". In: *Proceedings of the International Conference on Information Systems*. 2017.

[265] Erwan Moreau, Carl Vogel, and Marguerite Barry. "A paradigm for democratizing artificial intelligence research". In: *Innovations in Big Data Mining and Embedded Knowledge*. Springer, 2019, pp. 137–166.

[266] Jonathan T Morgan et al. "Tea and sympathy: crafting positive new user experiences on wikipedia". In: *Proceedings of CSCW*. ACM. 2013, pp. 839–848.

[267] Evgeny Morozov. *To save everything, click here: The folly of technological solutionism*. Public Affairs, 2013.

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

[269] Joe Moxley et al. "Writing analytics: Conceptualization of a multidisciplinary field". In: *Journal of Writing Analytics* 1 (2017).

[270] Jin Mu et al. "The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions". In: *International journal of computer-supported collaborative learning* 7.2 (2012), pp. 285–305.

[271] James Mullenbach et al. "Explainable prediction of medical codes from clinical text". In: *Proceedings of NAACL* (2018).

[272] Michael J Muller. "Participatory design: the third space in HCI". In: *The human-computer interaction handbook*. CRC press, 2007, pp. 1087–1108.

[273] Carol M Myford and Edward W Wolfe. "Detecting and measuring rater effects using many-facet Rasch measurement: Part I". In: *Journal of applied measurement* 4.4 (2003), pp. 386–422.

[274] Farah Nadeem et al. "Automated Essay Scoring with Discourse-Aware Neural Models". In: *Proceedings of BEA*. 2019, pp. 484–493.

[275] National Science Foundation. *Program on Fairness in Artificial Intelligence in Collaboration with Amazon*. Accessed 2020-07-25. URL: https://bit.ly/2BvaTXo.

[276] NCTE. *NCTE Position Statement on Machine Scoring*. https://bit.ly/3dQHaVY. Accessed 2020-06-30. 2013. URL: https://bit.ly/3dQHaVY.

[277] Hwee Tou Ng et al. "The CoNLL-2014 shared task on grammatical error correction". In: *Proceedings of CONLL*. 2014, pp. 1–14.

[278] Dong Nguyen. "Comparing automatic and human evaluation of local explanations for text classification". In: *Proceedings of NAACL*. 2018, pp. 1069–1078.

[279] Huy V Nguyen and Diane J Litman. "Argument mining for improving the automated scoring of persuasive essays". In: *AAAI Conference on Artificial Intelligence*. 2018.

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

[281] Matthew J Nunes. "The five-paragraph essay: Its evolution and roots in theme-writing". In: *Rhetoric Review* 32.3 (2013), pp. 295–313.

[282] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[283] Amy Ogan. "Reframing classroom sensing: promise and peril". In: *interactions* 26.6 (2019), pp. 26–32.

[284] Amy Ogan et al. "Oh dear stacy!: social interaction, elaboration, and learning with teachable agents". In: *Proceedings of CHI*. ACM. 2012, pp. 39–48.

[285] Christina Ortmeier-Hooper and Kerry Anne Enright. *Mapping new territory: Toward an understanding of adolescent L2 writers and writing in US contexts*. 2011.

[286] Ellis B Page. "The imminence of... grading essays by computer". In: *The Phi Delta Kappan* 47.5 (1966), pp. 238–243.

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

[288] Sita G Patel et al. "The achievement gap among newcomer immigrant adolescents: Life stressors hinder Latina/o academic success". In: *Journal of Latinos and Education* 15.2 (2016), pp. 121–133.

[289] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. "Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), p. 137.

[290] Judea Pearl. *Causality: models, reasoning and inference*. Springer, 2000.

[291] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12 (2011), pp. 2825–2830.

[292] Bonny Norton Peirce. "Social identity, investment, and language learning". In: *TESOL quarterly* 29.1 (1995), pp. 9–31.

[293] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of EMNLP*. 2014, pp. 1532–1543.

[294] Les Perelman. "When "the state of the art" is counting words". In: *Assessing Writing* 21 (2014), pp. 104–111.

[295] Ignacio J Pérez et al. "On dynamic consensus processes in group decision making problems". In: *Information Sciences* 459 (2018), pp. 20–35.

[296] Matthew E Peters, Sebastian Ruder, and Noah A Smith. "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP-2019)*. 2019, pp. 7–14.

[297] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. "Flexible domain adaptation for automated essay scoring using correlated linear regression". In: *Proceedings of EMNLP*. 2015, pp. 431–439.

[298] Wolter Pieters. "Explanation and trust: what to tell the user in security and AI?" In: *Ethics and information technology* 13.1 (2011), pp. 53–64.

[299] Christopher Pincock. "A role for mathematics in the physical sciences". In: *Noûs* 41.2 (2007), pp. 253–275.

[300] Stephen Porter. *Applying for an IES Grant*. 2015. URL: `https://stephenporter.org/research-methods/applying-for-an-ies-grant/`.

[301] Shrimai Prabhumoye et al. "Style Transfer Through Back-Translation". In: *Proceedings of the Association for Computational Linguistics*. 2018, pp. 866–876.

[302] Margaret Price. "Beyond" gotcha!": Situating plagiarism in policy and pedagogy". In: *College Composition and Communication* (2002), pp. 88–115.

[303] Danish Pruthi et al. "Learning to Deceive with Attention-Based Explanations". In: *Proceedings of ACL*. 2019.

[304] Ella Rabinovich et al. "Personalized Machine Translation: Preserving Original Author Traits". In: *Proceedings of the European Chapter of the Association for Computational Linguistics*. 2017, pp. 1074–1084.

[305] Peter Railton. "A deductive-nomological model of probabilistic explanation". In: *Philosophy of Science* 45.2 (1978), pp. 206–226.

[306] Diana Raufelder, Sandra Scherber, and Megan A Wood. "The interplay between adolescents' perceptions of teacher-student relationships and their academic self-regulation: Does liking a specific teacher matter?" In: *Psychology in the Schools* 53.7 (2016), pp. 736–750.

[307] Barbara Read, Becky Francis, and Jocelyn Robson. "Gender,'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays". In: *Assessment & Evaluation in Higher Education* 30.3 (2005), pp. 241–260.

[308] Y Malini Reddy and Heidi Andrade. "A review of rubric use in higher education". In: *Assessment & evaluation in higher education* 35.4 (2010), pp. 435–448.

[309] Byron Reeves and Clifford Nass. *How people treat computers, television, and new media like real people and places*. 1996.

[310] Justin Reich et al. "Remote Learning Guidance From State Education Agencies During the COVID-19 Pandemic: A First Look". In: *EdArXiv* (2020). https://doi.org/10.35542/osf.io/437e2.

[311] Ali Reza Rezaei and Michael Lovorn. "Reliability and validity of rubrics for assessment through writing". In: *Assessing writing* 15.1 (2010), pp. 18–39.

[312] Collin Rice. "Idealized models, holistic distortions, and universality". In: *Synthese* 195.6 (2018), pp. 2795–2819.

[313] Collin Rice. "Models Don't Decompose That Way: A Holistic View of Idealized Models". In: *The British Journal for the Philosophy of Science* 70.1 (2017), pp. 179–208.

[314] Collin Rice. "Moving beyond causes: Optimality models and scientific explanation". In: *Noûs* 49.3 (2015), pp. 589–615.

[315] John R Rickford and Russell John Rickford. *Spoken soul: The story of black English*. Wiley New York, 2000.

[316] Eric Riedel et al. "Experimental evidence on the effectiveness of automated essay scoring in teacher education cases". In: *Journal of Educational Computing Research* 35.3 (2006), pp. 267–287.

[317] Christoph Riedl and Anita Williams Woolley. "Teams vs. crowds: A field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-based problem solving performance". In: *Academy of Management Discoveries* 3.4 (2017), pp. 382–403.

[318] Brian Riordan, Michael Flor, and Robert Pugh. "How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models". In: *Proceedings of BEA*. 2019, pp. 116–126.

[319] Steven Ritter et al. "Cognitive Tutor: Applied research in mathematics education". In: *Psychonomic bulletin & review* 14.2 (2007), pp. 249–255.

[320] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. "Language models and Automated Essay Scoring". In: *arXiv preprint arXiv:1909.09482* (2019).

[321] Louie F Rodrıguez. "Moving beyond test-prep pedagogy: Dialoguing with multicultural preservice teachers for a quality education". In: *Multicultural Perspectives* 15.3 (2013), pp. 133–140.

[322] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A Primer in BERTology: What we know about how BERT works". In: *arXiv preprint arXiv:2002.12327* (2020).

[323] Rod D Roscoe and Danielle S McNamara. "Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom." In: *Journal of Educational Psychology* 105.4 (2013), p. 1010.

[324] Carolyn Rosé et al. "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning". In: *International journal of computer-supported collaborative learning* 3.3 (2008), pp. 237–271.

[325] Lippi-Green Rosini et al. *English with an accent: Language, ideology, and discrimination in the United States*. Psychology Press, 1997.

[326] Philip J Ruce. "Anti-money laundering: The challenges of know your customer legislation for private bankers and the hidden benefits for relationship management (the bright side of knowing your customer)". In: *Banking LJ* 128 (2011), p. 548.

[327] Sebastian Ruder et al. "Transfer learning in natural language processing". In: *Proceedings of NAACL: Tutorials*. 2019, pp. 15–18.

[328] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[329] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019).

[330] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. "Building automated vandalism detection tools for Wikidata". In: *Proceedings of the International Conference on the World Wide Web*. 2017, pp. 1647–1654.

[331] Bror Saxberg. "Learning engineering: the art of applying learning science at scale". In: *Proceedings of the Fourth (2017) ACM Conference on Learning Scale*. ACM. 2017, pp. 1–1.

[332] Cassandra Scharber, Sara Dexter, and Eric Riedel. "Students' Experiences with an Automated Essay Scorer." In: *Journal of Technology, Learning, and Assessment* 7.1 (2008), n1.

[333] Caroline Scheiber et al. "Gender differences in achievement in a large, nationally representative sample of children and adolescents". In: *Psychology in the Schools* 52.4 (2015), pp. 335–348.

[334] Jodi Schneider, Bluma S Gelley, and Aaron Halfaker. "Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review". In: *Proceedings of the international symposium on open collaboration*. ACM. 2014, p. 26.

[335] Jodi Schneider, Alexandre Passant, and Stefan Decker. "Deletion discussions in Wikipedia: Decision factors and outcomes". In: *Proceedings of WikiSym*. ACM. 2012, p. 17.

[336] Jodi Schneider et al. "Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups". In: *Proceedings of CSCW*. ACM. 2013, pp. 1069–1080.

[337] Jon Seger and JW Stubblefield. "Theoretical Evolutionary Ecology". In: *Bulletin of Mathematical Biology* 4.58 (1996), pp. 813–814.

[338] Andrew D Selbst et al. "Fairness and abstraction in sociotechnical systems". In: *Proceedings of FAccT*. ACM. 2019, pp. 59–68.

[339] Neil Selwyn. "Exploring the 'digital disconnect'between net-savvy students and their schools". In: *Learning, Media and Technology* 31.1 (2006), pp. 5–17.

[340] Neil Selwyn. "Re-imagining 'Learning Analytics'... a case for starting again?" In: *The Internet and Higher Education* (2020), p. 100745.

[341] Neil Selwyn. "What's the Problem with Learning Analytics?" In: *Journal of Learning Analytics* 6.3 (2019), pp. 11–19.

[342] Neil Selwyn and Scott Bulfin. "Exploring school regulation of students' technology use–rules that are made to be broken?" In: *Educational Review* 68.3 (2016), pp. 274–290.

[343] Sofia Serrano and Noah A Smith. "Is Attention Interpretable?" In: *Proceedings of ACL*. 2019.

[344] Mark D Shermis. "State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration". In: *Assessing Writing* 20 (2014), pp. 53–76.

[345] Mark D Shermis, Cynthia Wilson Garvan, and Yanbo Diao. "The Impact of Automated Essay Scoring on Writing Outcomes". In: *Annual Meeting of the National Council on Measurement in Education (NCME)*. 2008.

[346] Mark D Shermis and Ben Hamner. "Contrasting state-of-the-art automated scoring of essays: Analysis". In: *Proceedings of NCME*. 2012, pp. 14–16.

[347] Mark Shermis and Jill Burstein. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

[348] Annika Silvervarg et al. "The effect of visual gender on abuse in conversation with ECAs". In: *International conference on intelligent virtual agents*. Springer. 2012, pp. 153–160.

[349] Sharon Slade and Paul Prinsloo. "Learning analytics: Ethical issues and dilemmas". In: *American Behavioral Scientist* 57.10 (2013), pp. 1510–1529.

[350] Mona Sloane. "Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice". In: *Weizenbaum Conference 2019 "Challenges of Digital Inequality — Digital Education, Digital Work, Digital Life"*. 2019.

[351] Leslie N Smith. "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay". In: *arXiv preprint arXiv:1803.09820* (2018).

[352] Parinaz Sobhani, Diana Inkpen, and Stan Matwin. "From argumentation mining to stance classification". In: *Proceedings of the Workshop on Argumentation Mining at NAACL*. 2015, pp. 67–77.

[353] National Center for Special Education Research. *Building Evidence: What Comes After an Efficacy Study?* 2016.

[354] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer-Verlag, 1983.

[355] Luke Stark and Anna Lauren Hoffmann. "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture". In: *Journal of Cultural Analytics* (May 2019).

[356] Garold Stasser and William Titus. "Pooling of unshared information in group decision making: Biased information sampling during discussion." In: *Journal of personality and social psychology* 48.6 (1985), p. 1467.

[357] Potter Stewart, *concurring*. "Jacobellis v Ohio". In: *United States Supreme Court* 378 (1964), p. 184.

[358] Andreas Stolcke et al. "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: *Computational linguistics* 26.3 (2000), pp. 339–373.

[359] Sarah Strohkorb Sebo et al. "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 2018.

[360] Jennifer Stromer-Galley and Peter Muhlberger. "Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy". In: *Political Communication* 26.2 (2009), pp. 173–192.

[361] Julia Strout, Ye Zhang, and Raymond Mooney. "Do Human Rationales Improve Machine Explanations?" In: *Proceedings of the BlackboxNLP Workshop at ACL*. 2019, pp. 56–62.

[362] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of ACL* (2019).

[363] Sandeep Subramanian et al. "Multiple-attribute text style transfer". In: *Age* 18.24 (), p. 65.

[364] Bongwon Suh et al. "The singularity is not near: slowing growth of Wikipedia". In: *Proceedings of WikiSym*. ACM. 2009, p. 8.

[365] Emily Sullivan. "Understanding from Machine Learning Models". In: *British Journal for the Philosophy of Science* (2019).

[366] April Sutton et al. "Who Gets Ahead and Who Falls Behind During the Transition to High School? Academic Performance at the Intersection of Race/Ethnicity and Gender". In: *Social problems* 65.2 (2018), pp. 154–173.

[367] Kaveh Taghipour and Hwee Tou Ng. "A neural approach to automated essay scoring". In: *Proceedings of EMNLP*. 2016, pp. 1882–1891.

[368] Dario Taraborelli and Giovanni Luca Ciampaglia. "Beyond notability. Collective deliberation on content inclusion in Wikipedia". In: *IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*. 2010, pp. 122–125.

[369] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "Bert rediscovers the classical nlp pipeline". In: *Proceedings of ACL*. 2019.

[370] Elise Thomas. *How to hack your face to dodge the rise of facial recognition tech |WIREDUK*. `https://www.wired.co.uk/article/avoid-facial-recognition-software`. (Accessed on 05/01/2020). 2019.

[371] Marion Tillema et al. "Relating self reports of writing behaviour and online task execution using a temporal model". In: *Metacognition and Learning* 6.3 (2011), pp. 229–253.

[372] Zeynep Tufekci. "Big questions for social media big data: Representativeness, validity and other methodological pitfalls". In: *Proceedings of ICWSM*. 2014.

[373] Jean Underwood and Attila Szabo. "Academic offences and e-learning: individual propensities in cheating". In: *British Journal of Educational Technology* 34.4 (2003), pp. 467–477.

[374] Sowmya Vajjala. "Automated assessment of non-native learner essays: Investigating the role of linguistic features". In: *International Journal of Artificial Intelligence in Education* 28.1 (2018), pp. 79–105.

[375] Bas C Van Fraassen. "The pragmatics of explanation". In: *American Philosophical Quarterly* 14.2 (1977), pp. 143–150.

[376] Bas C Van Fraassen. *The scientific image*. Oxford University Press, 1980.

[377] Wendy P Van Ginkel and Daan van Knippenberg. "Group information elaboration and group decision making: The role of shared task representations". In: *Organizational Behavior and Human Decision processes* 105.1 (2008), pp. 82–97.

[378] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of NeurIPS*. 2017, pp. 5998–6008.

[379] Michael Veale, Max Van Kleek, and Reuben Binns. "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making". In: *Proceedings of CHI*. 2018, pp. 1–14.

[380] Matthew A Vetter, Zachary J McDowell, and Mahala Stewart. "From opportunities to outcomes: the Wikipedia-based writing assignment". In: *Computers and composition* 52 (2019), pp. 53–64.

[381] Fernanda B Viegas et al. "Talk before you type: Coordination in Wikipedia". In: *Proceedings of HICSS*. IEEE. 2007, p. 78.

[382] Elaine Lin Wang et al. "eRevis (ing): Students' revision of text evidence use in an automated writing evaluation system". In: *Assessing Writing* (2020), p. 100449.

[383] Wei Wang, Feixiang He, and Qijun Zhao. "Facial ethnicity classification with deep convolutional neural networks". In: *Chinese Conference on Biometric Recognition*. Springer. 2016, pp. 176–185.

[384] John Warner. "The Ed Tech Garbage Hype Machine: Behind the Scenes". In: *Inside Higher Ed* (2014). Accessed 2019-09-24. URL: https://bit.ly/1w3NdwS.

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

[386] Audrey Watters et al. "The problem with'personalisation'". In: *Australian Educational Leader* 36.4 (2014), p. 55.

[387] Melanie R. Weaver. "Do students value feedback? Student perceptions of tutors' written responses". In: *Assessment & Evaluation in Higher Education* 31.3 (2006), pp. 379–394.

[388] Candace West and Don H Zimmerman. "Doing gender". In: *Gender & society* 1.2 (1987), pp. 125–151.

[389] Eric Westervelt. *Meet The Mind-Reading Robo Tutor In The Sky*. https://bit.ly/318Tj4b. NPR Morning Edition. Accessed 2020-08-01. 2015.

[390] Edward M White, Norbert Elliot, and Irvin Peckham. *Very like a whale: The assessment of writing programs*. University Press of Colorado, 2015.

[391] Carl Whithaus. "Always already: Automated essay scoring and grammar checkers in college writing courses". In: *Machine scoring of student essays: Truth and consequences* (2006), pp. 166–176.

[392] Bernard E Whitley, Amanda Bichlmeier Nelson, and Curtis J Jones. "Gender differences in cheating attitudes and classroom cheating behavior: A meta-analysis". In: *Sex Roles* 41.9-10 (1999), pp. 657–680.

[393] Sarah Wiegreffe and Yuval Pinter. "Attention is not not explanation". In: *Proceedings of EMNLP*. 2019.

[394] Norbert Wiener. *The human use of human beings: Cybernetics and society*. 320. Da Capo Press, 1988.

[395] David Wiley and John Levi Hilton III. "Defining OER-enabled pedagogy". In: *International Review of Research in Open and Distributed Learning* 19.4 (2018).

[396] Andrew Wilkins. "Push and pull in the classroom: competition, gender and the neoliberal subject". In: *Gender and Education* 24.7 (2012), pp. 765–781.

[397] Ben Williamson. *Code Acts in Education: Re-Engineering Education*. 2020. URL: https://nepc.colorado.edu/blog/re-engineering-education.

[398] Ben Williamson. "Decoding ClassDojo: psycho-policy, social-emotional learning and persuasive educational technologies". In: *Learning, Media and Technology* 42.4 (2017), pp. 440–453.

[399] David M Williamson, Xiaoming Xi, and F Jay Breyer. "A framework for evaluation and use of automated scoring". In: *Educational measurement: issues and practice* 31.1 (2012), pp. 2–13.

[400] Joshua Wilson and Amanda Czik. "Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality". In: *Computers & Education* 100 (2016), pp. 94–109.

[401] Joshua Wilson and Rod D Roscoe. "Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy". In: *Journal of Educational Computing Research* (2019), p. 0735633119830764.

[402] Kenneth G Wilson. "Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture". In: *Physical review B* 4.9 (1971), p. 3174.

[403] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis". In: *Proceedings of EMNLP*. 2005.

[404] Terry Winograd, Fernando Flores, and Fernando F Flores. *Understanding computers and cognition: A new foundation for design*. Intellect Books, 1986.

[405] Bronwyn Woods et al. "Formative essay feedback using predictive scoring models". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 2071–2080.

[406] James Woodward. "Capacities and invariance". In: *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grunbaum* (1994), p. 283.

[407] James Woodward. "Explanation, invariance, and intervention". In: *Philosophy of Science* 64 (1997), S26–S41.

[408] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[409] James Woodward. "Scientific Explanation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University, 2017.

[410] Bo Xiao and Izak Benbasat. "E-commerce product recommendation agents: use, characteristics, and impact". In: *MIS quarterly* 31.1 (2007).

[411] Lu Xiao and Nicole Askin. "What influences online deliberation? A Wikipedia study". In: *Journal of the Association for Information Science and Technology* 65.5 (2014), pp. 898–910.

[412] Lu Xiao and Jeffrey Nickerson. "Imperatives in Past Online Discussions: Another Helpful Source for Community Newcomers?" In: *Proceedings of HICSS*. 2019.

[413] Lu Xiao and Niraj Sitaula. "Sentiments in Wikipedia Articles for Deletion Discussions". In: *International Conference on Information*. Springer. 2018, pp. 81–86.

[414] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *Proceedings of the International Conference on Machine Learning*. 2015, pp. 2048–2057.

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

[416] Diyi Yang. "Computational Social Roles". PhD thesis. Carnegie Mellon University, 2019.

[417] Diyi Yang et al. "Identifying semantic edit intentions from revisions in wikipedia". In: *Proceedings of EMNLP*. 2017, pp. 2000–2010.

[418] Diyi Yang et al. "Let's Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms". In: *Proceedings of NAACL*. 2019, pp. 3620–3630.

[419] Diyi Yang et al. "Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities". In: *Proceedings of CHI*. ACM. 2019, p. 344.

[420] Diyi Yang et al. "Who Did What: Editor Role Identification in Wikipedia." In: *ICWSM*. 2016, pp. 446–455.

[421] Qian Yang, Nikola Banovic, and John Zimmerman. "Mapping machine learning advances from hci research to reveal starting places for design innovation". In: *Proceedings of CHI*. 2018, pp. 1–11.

[422] Zhilin Yang et al. "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". In: *Proceedings of EMNLP*. 2018, pp. 2369–2380.

[423] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Proceedings of NeurIPS*. 2019, pp. 5753–5763.

[424] Helen Yannakoudakis and Ted Briscoe. "Modeling coherence in ESOL learner texts". In: *Proceedings of BEA*. Association for Computational Linguistics. 2012, pp. 33–43.

[425] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. "Understanding the effect of accuracy on trust in machine learning models". In: *Proceedings of CHI*. 2019, pp. 1–12.

[426] Nicholas Yoder. "Teaching the Whole Child: Instructional Practices That Support Social-Emotional Learning in Three Teacher Evaluation Frameworks. Research-to-Practice Brief". In: *Center on Great Teachers and Leaders* (2014).

[427] Iris Marion Young. *Justice and the Politics of Difference*. Princeton University Press, 1990.

[428] John Zerilli et al. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" In: *Philosophy & Technology* (2018), pp. 1–23.

[429] Torsten Zesch and Oren Melamud. "Automatic generation of challenging distractors using context-sensitive inference rules". In: *Proceedings of BEA*. 2014, pp. 143–148.

[430] Justine Zhang et al. "Characterizing online public discussions through patterns of participant interactions". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–27.

[431] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. "Fine-grained sentiment analysis with faithful attention". In: *arXiv preprint arXiv:1908.06870* (2019).