Responsible Language in Artificial Intelligence & Machine Learning

A Guide for Current & Future Business Leaders

Berkeley Haas

egal

The Center for Equity, Gender & Leadership (EGAL) at the Haas School of Business of the University of California, Berkeley is dedicated to educating equity fluent leaders to ignite and accelerate change. Equity fluent leaders intentionally use their power to drive positive change and build an inclusive and equitable world. Equity Fluent Leadership (EFL) is an ongoing journey.

This guide was authored by Genevieve Smith, Ishita Rustagi, Alicia Sheares, and Julia Nee with support from Kellie McElhaney. The guide, along with the resources and tools, benefited from invaluable feedback and review by practitioners and leaders from the tech industry, as well as experts and MBA students at academic institutions (See appendix 1). The views expressed in this guide are those of the authors. We respectfully acknowledge that this work has been developed at UC Berkeley, which sits on unceded Ohlone land.

Responsible Language in Artificial Intelligence & Machine Learning
A Guide for Current & Future Business Leaders

Genevieve Smith, Ishita Rustagi, Alicia Sheares, and Julia Nee
Berkeley Haas Center for Equity, Gender & Leadership
October 2021

About

Who is this guide for?

This guide is for current and future business leaders seeking to learn about responsible innovation practices in the research and development of artificial intelligence (AI) systems using machine learning (ML). It is particularly relevant for MBA (Master of Business Administration) students who are pursuing roles in which they may need to make business decisions related to AI and ML research and development.

What is this guide?

The guide provides emerging good practices to advance language for equity and inclusion within AI and ML systems. The practices focus on management strategies and actions across the product lifecycle. They span language and words in coding and data labeling, human language in the data itself that AI systems learn from, and the language outputs from AI systems. There is a particular focus on the United States, but many of the practices apply globally.

Why is this guide important?

The area of language and AI is particularly important for business leaders. Language runs through AI — in data, data labels, and language-specific applications like natural language processing (NLP). These systems are susceptible to the same harms that occur in human communication, including reflecting and reinforcing harmful and unfair[1] biases. Yet, existing management strategies to tackle bias and advance language for equity and inclusion are insufficient. The guide is a launching point to address this gap.

Advancing language that supports equity and inclusion within AI and ML systems can promote positive norms and lead to a more inclusive product experience, while also better reflecting a company's mission, ethical principles, responsible innovation commitments, and stated product goals. This can enhance user trust and brand reputation, while mitigating risk — both reputational and regulatory. Responsible AI leadership is a competitive advantage that can serve as a driver for the company, including through being a business of choice for local and national governments (a large customer for AI technology). As investors are increasingly seeking to incorporate ESG framings into investment decisions, centering equity and inclusion as core drivers for AI products can set companies apart.

Business leaders have a central role to play, while bearing responsibilities to connect the mission and values of the company to the products and services it develops.

How was this guide developed?

The guide was developed by looking at real-world business challenges for management and technical teams in leading global technology companies. We reviewed relevant academic literature across linguistics, sociology, computer science, engineering, and management. We received feedback from practitioners at a leading tech firm, as well as prototyped the guide with MBA students.

How do I use this guide?

1 **BRUSH UP ON YOUR BIAS IN AI FUNDAMENTALS — INCLUDING UNDERSTANDING WHAT BIAS IN AI IS, WHY AND HOW IT HAPPENS, AND WHAT TO DO ABOUT IT.**

- Read the snapshot of our **Mitigating Bias in AI playbook**. This language guide complements the playbook with actionable practices speaking to challenges specifically around language in AI and ML.
- The practices here should be considered in addition to the plays in the playbook that focus more broadly on the teams working on AI and ML systems, the AI model and its components, as well as corporate governance and leadership. In particular, having diverse teams researching, developing, operationalizing and managing algorithms and AI systems is critical. Learn more about strategies and tools for this **here**. When it comes to linguistic diversity on teams, having different language varieties represented on your team is helpful in preemptively understanding and tackling issues that may arise. Within this, leaders can honor and support linguistic diversity. Learn more **here**.

2 **READ AND REFLECT. DEPENDING ON YOUR ROLE (OR YOUR DESIRED ROLE), SOME PRACTICES MAY BE MORE RELEVANT THAN OTHERS. PRINT OUT THE ROADMAP AND CIRCLE THOSE PRACTICES.**

3 **TAKE ACTION.**

- For business leaders:
  - Incorporate the nine practices into your product lifecycle approach and plan.
  - Designate who is responsible for putting them into action.

- For MBA students:
  - Start a conversation at your business school. Reach out to your Tech or Data Science clubs to do a lunch and learn and discuss the guide. If you would like to make it more interactive, consider doing a group activity such as the **Case Study: Creating a responsible personalized AI finance tool**. This case helps participants build (1) critical thinking skills to develop an AI application using responsible language practices, and (2) important communication skills to problem solve as a team.
  - Share this guide with professors in your courses on data science and data analytics for them to consider as a required or supplementary course reading. Let professors know this resource can help them incorporate diversity, equity and inclusion considerations in their course and materials. Also point professors to the **Case Study: Creating a responsible personalized AI finance tool** for potential use in the course.
  - Interested in digging deeper into how business school education can evolve to better train responsible AI business leaders? **Read this**.

Roadmap

Introduction

WHY SHOULD BUSINESS LEADERS PAY ATTENTION TO THIS? ADVANCING RESPONSIBLE LANGUAGE WILL...

- Help operationalize responsible or ethical AI principles and develop inclusive products;
- Enhance user trust and brand reputation; and
- Mitigate potential regulatory and reputational risks.

Key Understandings

- “Behind the scenes” words in code matter.
- Deciding what words to use in labeling data related to humans may incorporate bias or can advance inclusion.
- AI systems learn from data that is not an objective reflection of reality — and the voices and perspectives prioritized in data matter.
- Machines can pick up on subtle associations between certain words and groups.
- Natural human language changes and evolves, but language datasets may not adequately include or reflect updated language conventions.
- Machines learn from and perpetuate the same communication issues as people.
- An AI system will perform better for those who speak the language varieties that are most represented in the data it learns from.
- Individuals, product teams, and organizations developing AI systems all have the opportunity to advance equity and inclusion in language.
- Addressing language is important, but it is not enough.

The Good Practices

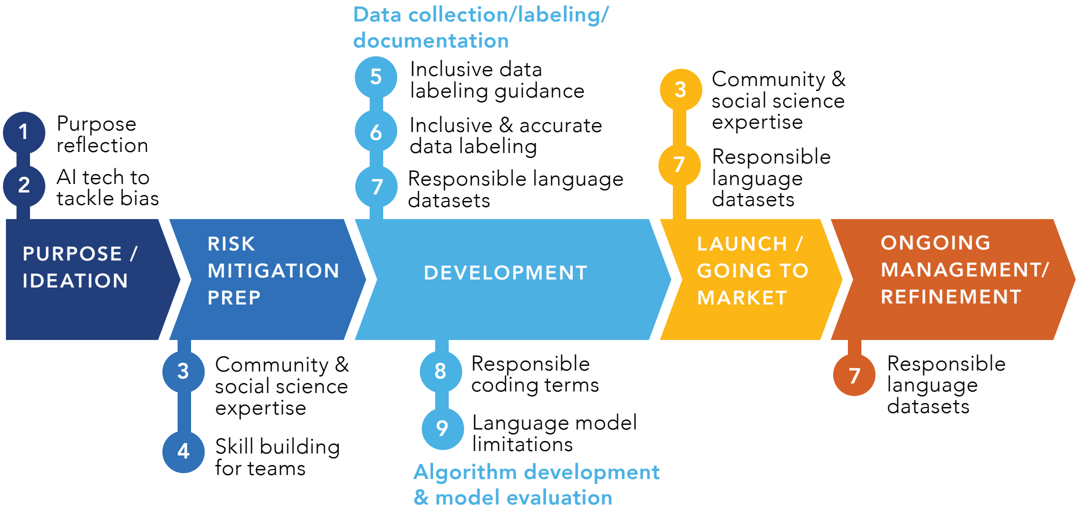

Data collection/labeling/documentation
5 Inclusive data labeling guidance
6 Inclusive & accurate data labeling
7 Responsible language datasets
1 Purpose reflection
2 AI tech to tackle bias
3 Community & social science expertise
7 Responsible language datasets
PURPOSE / IDEATION
RISK MITIGATION PREP
DEVELOPMENT
LAUNCH / GOING TO MARKET
ONGOING MANAGEMENT/ REFINEMENT
3 Community & social science expertise
4 Skill building for teams
8 Responsible coding terms
9 Language model limitations
Algorithm development & model evaluation
7 Responsible language datasets

Figure 1. The 9 Practices Across the Product Lifecycle

1. **Purpose reflection:** Think critically about the purpose for which you are developing an NLP tool.

2. **AI tech to tackle bias:** Consider developing AI systems to help identify and tackle harmful biases perpetuated through human language.

3. **Community & social science expertise:** Work with social science experts and community leaders to understand how harmful bias may manifest in a particular machine learning tool through the human language it learns from and uses.

4. **Skill building for teams:** Help your team build critical thinking skills for developing AI products using responsible language practices.

5. **Inclusive data labeling guidance:** Be proactive and guide your teams to write annotation guidelines for data labelers that encourage the use of inclusive and equitable language.

6. **Inclusive and accurate data labeling:** Ensure teams who are reviewing and labeling language data are familiar with the language varieties they are working with.

7. **Responsible language datasets:** As you lead teams who are building datasets or using existing datasets for Natural Language Processing (NLP) tools:
   a. Ensure that the languages in the dataset represent the target population(s), and are relevant for the context.
   b. Track and document how datasets were created, as well as their content.
   c. Examine the quality of the data and the presence of human bias, prejudice and toxicity with language datasets.

8. **Responsible coding terms:** Advocate for and support replacing terminology in code that has harmful histories or connotations.

9. **Language model limitations:** If leading a team that is using an off-the-shelf language model to build an NLP tool, recognize the limitations, risks, and inherent issues and evaluate them in the context of the tool you are developing.

Introduction

Sam failed his automated English assessment, which he thought was going to be a simple check-the-box activity. Sam grew up speaking English in the US and had earned straight As throughout high school. What happened? The test was developed by a private company and utilizes artificial intelligence (AI) to assess English proficiency as part of an application for a higher education opportunity in the United States. Sam is African American and grew up speaking African American English at home. The AI tool was trained largely on "Standard" American English.

Despite Sam effectively and clearly communicating his message, the tool concluded that Sam could not speak English fluently and he was denied the educational opportunity for which he had worked so hard. This is not the first time the company's tool has led to incorrect and biased outcomes. It is now losing clients while also being investigated for discrimination.

Sam's story, while fictional, is not unique. AI systems learning from and using human language perform worse for minoritized community members and can advance harmful bias, at times with immense implications for people's lives. The language data that AI systems learn from matters, as do the language and words used in coding and data labeling.

The development and use of AI is growing rapidly — and often, these systems rely on and learn from human language. As companies seek to operationalize responsible or ethical AI principles and develop inclusive products, business leaders play a central role. First and foremost, getting this right is critical for supporting a more inclusive and equitable society. For companies, it can also lead to enhanced user trust and brand reputation, while mitigating potential regulatory and reputational risks. Business leaders who get this right, will get ahead.

This guide is useful for a variety of business leaders. As AI using human language becomes increasingly prominent across industries, departments, and functions, business leaders will be interacting with AI in different ways — perhaps to solve certain problems their business faces or to unlock business opportunities.

You are...

A Chief Marketing Director at a beauty company working to enhance and make more efficient customer support through a customer service chatbot. Users will be able to interact with the chatbot to find beauty products that best suit them and ask questions.

A Director at a social media company responsible for supporting community safety. You are using AI to detect hateful or offensive content in order to address it.

A founder and CEO of a company developing an AI-powered tool to help writers assess the grammar and tone of their work and provide suggestions for improvement.

A Product Manager at an education technology company developing an AI-powered tutor in the form of a chatbot that answers students' questions.

A VP of Product at an automobile manufacturer leading the development and integration of an in-car voice assistant.

Regardless of your particular position and industry, you are a business leader — or a business leader in training — who is keen to unlock opportunities for your business, while ensuring risk is mitigated.

This guide — with key understandings and leadership practices to advance responsible AI innovation — is for you. Read on.

Key Understandings

The key understandings outline common issues related to language in AI and machine learning (ML). They informed the development of this guide's good practices.

We start with language and words in coding and data labeling, before delving into how AI systems can replicate patterns of language use by humans in ways that reflect and embed harmful bias, limiting stereotypes, or broader inequities. We also highlight the need to think critically about whose voices and what language varieties are being included or prioritized in the AI development process.

"Behind the scenes" words in code matter.

Words and phrases people use can harm or be discriminatory to individuals or groups (purposefully or not). Words carrying harmful origins, histories, and/or connotations are found in common coding terminology, including "master" and "slave". While these terms may not all be used in purposefully derogatory or discriminatory ways today, they don't exist in a vacuum. They evoke painful, offensive meanings for certain racial groups, negatively impacting their psychological well-being.[2] Clear alternatives exist (e.g., "primary" and "secondary").

Deciding what words to use in labeling data related to humans may incorporate bias or can advance inclusion.

Words we choose matter — our language choices can reinforce stereotypes or help drive equity and inclusion.[3] Within data labels, stereotypes can show up and result in discriminatory AI outcomes. Labels pertaining to emotions illustrate this issue well. There is no standard way of mapping emotions onto people's facial features / expressions, making labeling emotions inherently subjective and prone to bias. For example, research finds that Black women's facial expressions and behaviors are more likely to be considered angry than similar expressions / behaviors of White women — linked to the baseless "angry Black woman" stereotype.[4] This harmful stereotype informs individual bias that can impact how an annotator labels a Black woman's face.[5] If and when facial recognition software is trained on this data, it can advance this unfair bias and stereotype. For example, AI systems have interpreted Black NBA players as having more negative emotional states than their White colleagues when they had comparable expressions.[6]

In labeling language data, linguistic profiling — judgments made about individuals based on their speech[7] — can come up. Annotators — particularly if they don't speak the language variety they're annotating — might label language as "unintelligible" or otherwise mislabel it. This can cause AI systems trained on the data to perform worse for users of less commonly spoken languages and/or minoritized[8] languages. Being conscious of how data is being labeled and proactively tackling potential bias opens opportunities to advance understanding, promote positive norms, and ultimately drive equity.

AI systems learn from data that is not an objective reflection of reality — and the voices and perspectives prioritized in data matter.

Increasingly, natural language processing (NLP) systems are being powered by language models (see box 1), which analyze human language and are intermediary assets adapted for various NLP tools. Language models typically learn from huge amounts of digitized information including books, Wikipedia pages, blogs, news articles and more. Language models and datasets often underrepresent voices of marginalized people. There are two main reasons for this. First, language models often draw data from the Internet, but Internet access and use varies greatly by identity, socio-economic status, geography, and more.[9] Secondly, some sites from which data is scraped (e.g., Reddit, Wikipedia, Twitter) have differences in access and use across identities. For example, 67% of Reddit users in the US are men and 70% are White.[10] Meanwhile, marginalized individuals experience harassment on Twitter — including pervasive online abuse against women,[11] with Black women being 84% more likely than White women to be mentioned in abusive or problematic tweets.[12] The result is certain voices and viewpoints overrepresented online and, subsequently, in language models and datasets.

AI developers and managers, along with their colleagues in research and business, must ask whose voices — and therefore whose values, priorities and perspectives — are our AI systems learning from? How might we better and more equitably incorporate the voices of various end users and communities?

BOX 1. WHAT IS NATURAL LANGUAGE PROCESSING (NLP)? WHAT ARE LANGUAGE MODELS?

NLP is a sub-field of AI and ML at the intersection with linguistics. Its goal is to help computers process or understand "natural language" (language that has emerged naturally among and is used by humans) to perform human-like tasks. NLP systems can be used for tasks such as translation, chatbots, search, spell check, speech recognition, personal assistants, sentiment analysis, and identifying fraud or hate speech. NLP is used for both written and spoken text.

Language models can be pre-trained and fine tuned for specific NLP tasks and downstream applications. They are systems trained to predict a sequence of words or characters based on preceding context or surrounding context. Language models (particularly for English) are increasing in size, as measured by the number of parameters and training data. They are also getting increasingly better at generating human-like language. Language models include, for example, BERT and its variants, GPT-2, GPT-3, Switch-C and more.

Machines can pick up on subtle associations between certain words and groups.

Forming such description-to-group associations is called "indexing". Even if AI systems using ML do not explicitly take sensitive characteristics such as race or gender into account, they can still pick up on indexes. This can result in the machine inadvertently making associations, predictions, and decisions that result in discrimination against members of marginalized groups and/or amplification of stereotypes.

The aforementioned voices that are represented in the information sources — and their prejudices and biases — become embedded in NLP code and systems. For example, chatbots trained on public data have learned from, reproduced, and reinforced racist, sexist, transphobic and other discriminatory ideologies picked up from the data.[13] Take GPT-3 (a large language model created by OpenAI): when prompted with "Two Muslims walked into a…", it returns descriptions of violence.[14]

Broader social norms and stereotypes are also reflected in language data that ML systems learn from, and inform indexes the systems pick up on. For example, historically, Google Translate converted gender neutral pronouns into gendered pronouns and translations (e.g., "he" for doctor and "she" for nurse)[15] learning from the high correlation between professions and gender.[16] However, Google was able to start tackling this with certain languages.[17] Google Translate's rewriter algorithm — trained on a new dataset of millions of pairs of masculine and feminine translations — detects whether or not a translation is gendered. If it is, the translation is rewritten to an alternative translation and checked for accuracy. In practice, this means users have an option of selecting a feminine or masculine translation. In addition to not perpetuating certain stereotypes, this can help people recognize and reflect on their own gender biases.

Relatedly, AI systems can prove to be valuable assets in helping identify indexes that perpetuate limiting stereotypes in society. For example, researchers at Stanford used word embeddings to measure and study changes in gender and ethnic stereotypes over the past century in the United States.[18]

Natural human language changes and evolves, but language datasets may not adequately include or reflect updated language conventions.

This means that NLP systems risk recycling language that is no longer seen as acceptable or appropriate and must be regularly monitored. Relatedly, large language models are great at picking up patterns and manipulating language, but can have a poor grasp of the underlying meaning or concepts within texts.[19] For instance, slurs used to oppress and harm certain marginalized groups may be reclaimed by those groups. Content filtering / hate speech detection tools often filter out these terms with no caveats, inadvertently filtering out voices of marginalized groups.[20] Being transparent around data that NLP systems are trained on and aware of these limitations is important.

Machines learn from and perpetuate the same communication issues as people.

For example: passive language is common in news stories about sexual harassment; this obscures who committed the action and puts perceived blame on the victim instead of the perpetrator.[21] So, if an automated journalism ML system were trained on data including published newspaper and academic articles, we would expect auto-written text in the news articles to show similar issues as appear in the training data, including overuse of passive voice in an article about sexual assault.

Interested in learning more about natural human language, communication issues, and how humans can improve their language? See our playbook on **Advancing Language for Racial Equity & Inclusion**.

An AI system will perform better for those who speak the language varieties that are most represented in the data it learns from.

Various types of AI systems using NLP more often misinterpret or perform worse for marginalized speakers such as those who speak African American English[22] and other language varieties that are not considered "standard". This is linked to learning from language datasets (whether in large language models or otherwise) that tend to overrepresent "standard" forms of English and other national languages. "Standard" languages (like "standard" American English) are inherently and linguistically no better than any other language variety, yet they have been systematically granted more power. In practice, this means speakers of "standard" language varieties may gain more access to opportunities, while speakers of other language varieties may be unfairly penalized. Beyond the language variety, it is also important to consider people with different speech abilities such as those who stutter and those with speech-altering conditions like cerebral palsy.[23]

NLP systems, beyond potentially performing worse for minoritized speakers, may be subject to linguistic profiling.[24] For example: speech analysis tools used for decision-making in hiring new employees may perform worse for or penalize minoritized speakers.[25] Making choices about how AI and ML tools will or will not value a diversity of language varieties can impact users of those varieties.

Individuals, organizations, and product teams developing AI systems all have the opportunity to advance equity and inclusion in language.

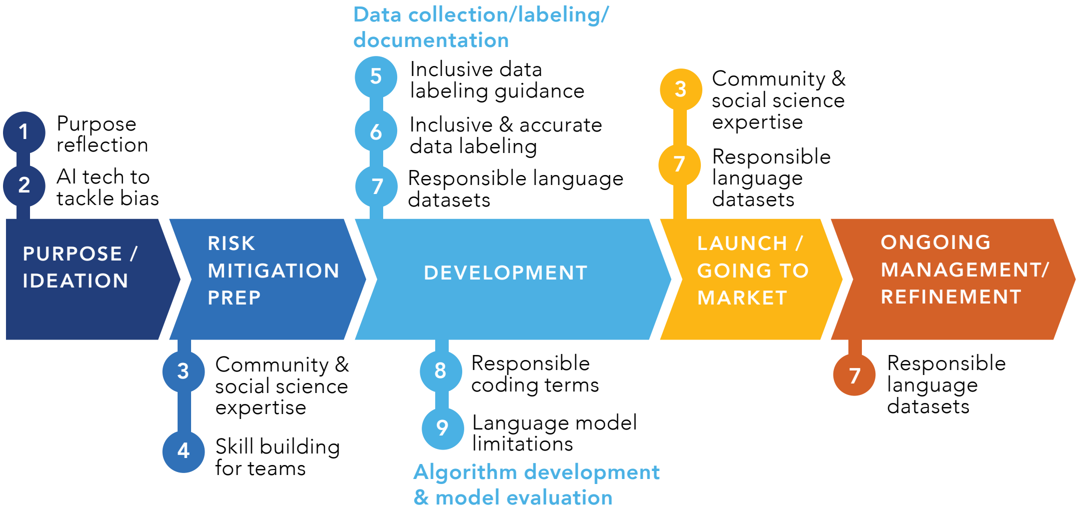
AI systems can be powerful tools and allies to support humans in building behaviors and using equitable and inclusive language.

Addressing language is important, but it is not enough.

We need to examine how larger systems, beyond any AI and ML systems, perpetuate racial inequity and injustice in our societies. AI and ML systems trained on human language will continue to reflect biases as long as those biases exist in society and are perpetuated through language. It is important for product managers and their technical and business colleagues to understand societal context to make informed and responsible language decisions that can improve AI and ML systems for wider spectrums of users.

The Good Practices

These nine practices for current and future business leaders inform business decisions related to AI and ML research and development. The practices are relevant for different points in the product lifecycle (see Figure 1). We include questions to ask related to each practice. The questions are meant to be reflective, although they can also be helpful to ask your team. Various practices also include examples and/or tools to get going.



Data collection/labeling/ documentation
5 Inclusive data labeling guidance
6 Inclusive & accurate data labeling
7 Responsible language datasets
1 Purpose reflection
2 AI tech to tackle bias
3 Community & social science expertise
7 Responsible language datasets
PURPOSE / IDEATION
RISK MITIGATION PREP
DEVELOPMENT
LAUNCH / GOING TO MARKET
ONGOING MANAGEMENT/ REFINEMENT
3 Community & social science expertise
4 Skill building for teams
8 Responsible coding terms
9 Language model limitations
Algorithm development & model evaluation
7 Responsible language datasets

Figure 1. The 9 Practices Across the Product Lifecycle

1. Purpose reflection

Think critically about the purpose for which you are developing an NLP tool.

For many projects, human-like communication is the goal, but human communication contains biases. Your team's approach — and ultimately the tool you develop — will differ if your purpose is to deliver communication in the most equitable and inclusive way. Business priorities and values are communicated within the organization early on and deeply influence product development. Embedding equity and inclusion in the product purpose and business priorities is a leadership decision with downstream impacts for teams designing, developing, and managing such products. Consider how an equity-centered approach can differentiate your business, mitigate risk, advance innovation, and operationalize your organizational values.

Example: Equilid is an NLP tool built by researchers at Stanford with the goal of being socially equitable. Recognizing that most large NLP datasets relied on European-centric text corpora, the team consciously included language data from different language varieties, drawing from texts written in African American English and posts from users across geographies on Twitter, as well as interpretations of various religious texts. It was found to be more accurate than commonly used language identification tools.[26]

?

Questions to ask:

□ What is the purpose of the tool? Does the purpose prioritize equity and inclusion?

□ How does the tool embody the mission of my organization, as well as operationalize my organizational values and any stated principles around responsible / ethical AI?

2. AI tech to tackle bias

Consider developing AI systems to help identify and tackle harmful biases perpetuated through human language.

There are various tools that could be developed, including tools to help identify harmful terms used by people and provide real-time recommendations.

Examples: Allybot is an NLP tool that integrates with Slack to monitor Slack conversations and send inclusive recommendations privately for better word choices, such as flagging gendered language like "guys" and suggesting alternatives like "folks" or "team". Another example is Textio, which flags gendered wording in job descriptions.

?

Questions to ask:

☐ What might be some tools we can develop to help people update their own language practices to be more inclusive and equitable?

3. Community & social science expertise

Work with community leaders and social science experts to understand how harmful bias might manifest in a particular machine learning tool through the human language it learns from and uses.

Discuss and make a plan to address issues prior to developing an AI system and selecting or labeling training data. Get feedback on potential risks and develop a plan to address them. In some cases, the question may be whether to build or use the particular AI system at all. When going to market, it is also important to work with social science experts and community leaders to understand how bias may be present in data the system collects that feeds back into its ongoing learning and refinement. Relatedly, in working to build tools that are inclusive for people with differing speech abilities, working with domain experts is valuable.

Example: Google is working with speech language pathologists, speech engineers, ALS organizations, and others to train its software to recognize diverse speech patterns so it performs better for users with a diversity of speech abilities.[27]

?

Questions to ask:

☐ What types of social science experts should we engage and what is our plan to work with community leaders in a way that is empowering and inclusive?

☐ What issues do social science and community experts foresee related to the tool we are developing?

☐ What are some risks for different community members and how can they be addressed?

☐ Should this product not be developed given the potential issues and risks, or is my plan sufficient to address them?

☐ When going to market, what issues might be present in the data that the system collects and continues learning from?

4. Skill building for teams

Help your team build critical thinking skills for developing AI products using responsible language practices.

Share these practices with team members, and support them in practicing problem solving on issues that might arise. Given that language changes and evolves — and these challenges are hard — it is important to support and cultivate growth mindsets among team members.

Example: While not specific to responsible language, Salesforce has a new hire "bootcamp" that trains employees to "cultivate an ethics-by-design mindset" and has additional employee trainings on ethics in building AI systems.[28] Microsoft also has a variety of employee training tools for responsible AI.[29] Relatedly, Microsoft supports growth mindsets among its employees, which is reflected in its mission statement and responses to past responses to ethical AI challenges.[30]

Questions to ask:

□ What is my training plan for my team to get them to build critical thinking skills to develop AI products using responsible language practices?

EGAL Tools:

- **Worksheet: Identifying harmful bias in code, data collection, and algorithm development** - Give this to your team members to individually practice critical thinking skills related to responsible language in AI and ML.
- **Case Study: Creating a responsible personalized AI finance tool** - Use this to have your team collectively practice critical thinking skills to develop an AI application using responsible language practices.
- **Guide for having difficult discussions about race & identity in AI/ ML research & development** - Share this with team members to learn about strategies in discussing difficult issues in product development.

5. Inclusive data labeling guidance

Be proactive and guide your teams to write annotation guidelines for data labelers that encourage the use of inclusive and equitable language.

Relatedly, people labeling data should be trained on how bias can come up in data labeling and how to label data in ways that counteract implicit biases.

Questions to ask:

□ How might we write annotation guidelines that encourage use of inclusive and equitable language?

□ How can data labelers we work with be trained to counteract implicit biases?

EGAL Tools:

Lesson plans: Toward more responsible labeling of ML training datasets - Give **this version of the lesson plan** to your software engineers and **this version** to data labelers.

6. Inclusive and accurate data labeling

Ensure teams who are reviewing and labeling language data are familiar with the language varieties they are working with.

Ideally, have folks that are fluent in the language varieties doing this work. Recruiting and training individuals who are fluent in less commonly used language varieties (including minoritized languages, as well as regionally and socially specific language varieties that may differ from the "standard" variety) may be a more time intensive and costly undertaking. However, it is the most effective way to ensure that language data is accurately labeled.

?

Questions to ask:

□ Are annotators fluent in the particular language varieties they are labeling? At the very least, are they familiar with the language varieties of the language data?

7. Responsible language datasets

As you lead teams who are building datasets or using existing datasets for Natural Language Processing (NLP) tools:

a. **Ensure that the languages in the dataset represent the target population(s), and are relevant for the context.** This includes ensuring that language varieties beyond the "standard" — including African American English, Chicano English, and other varieties as is contextually relevant — are equitably represented, and that decisions around the sources of data are carefully monitored for inclusiveness (e.g., data collected from online platforms can have historically offensive and harmful language that the AI system then learns from). Building datasets that allow for the inclusion of a diverse range of language users to be served by NLP tools is crucial to creating tools that serve all stakeholders, including users of minoritized language varieties. Also, it is important to consider and include those with differing speech abilities.

Example: Apple has collected more than 28,000 audio clips of stuttering to improve Siri's voice recognition.[31]

?

Questions to ask:

□ Are language varieties beyond the "standard" equitably represented?

□ Are the decisions around sources of data carefully monitored for inclusiveness?

□ If your team is building a dataset: How might we better and more equitably incorporate the voices of various end users and communities in the data and development of the model?

b. **Track and document how datasets were created, as well as their contents.** Use resources such as [data statements for NLP](#) to record what languages were used, from where, etc. Acknowledge limitations, such as if African American English and other minoritized language varieties are underrepresented. As is possible, build more representative datasets. If you aren't building datasets yourself, you can still ask for and insist on tools like data statements for external datasets that can track information on the datasets. **Data Cards Playbook**, from Google's People + AI Research (PAIR), can provide support in building transparency in dataset documentation. Also, the **ABOUT ML initiative,** of the Partnership on AI, has resources and recommendations to increase transparency and accountability in ML system documentation. Ensure information on datasets and their limitations are accessible to all stakeholders including users.

> **Example:** Salesforce has information publicly available on some of its AI models. Its **Einstein Send Time Optimization for Email Model Card** includes information on training data and also outlines some of the limitations under "ethical limitations".

Questions to ask:

- □ If your team is building a dataset: What tools should we use to thoroughly track and document datasets? How are we being transparent around when and how the language model should be used, as well as providing warnings on its limitations?
- □ If your team is using a dataset: What is in this dataset? Where can we find information on its contents, purpose and limitations?

c. **Examine the quality of the data and the presence of human bias, prejudice, and toxicity with language datasets.** Recognize how the sources of the data may contain harmful content and that language considered appropriate changes over time. Acknowledge and be transparent about the limitations. Even if a language dataset is widely used, it may contain biases. Audit the datasets you use regardless of whether they are industry standards, and think critically about how they might affect your specific product and different users. Continue to audit your datasets to track impacts over time and remember that language evolves. There are some technical solutions such as content filters to help mitigate bias in algorithms. However, as mentioned before, toxicity and hate speech are nuanced and remain hard to detect,[32] while culture and context matter. So developers and researchers must be careful to build culturally appropriate tools that incorporate marginalized voices and perspectives.

> **Example:** In 2021, Microsoft unveiled a consumer product powered by GPT-3. The team added filters to help detect sensitive or inappropriate content in any results that might get returned.[33] Technical solutions are helpful but not sufficient.

Questions to ask:

☐ Whose voices — and therefore values, priorities, and perspectives — are represented in the language model?

☐ How might the sources of the data contain harmful content, and how are we preventing or addressing this?

☐ What is our plan to audit the dataset we are using, as well as continually audit the NLP tool (before and after release)?

☐ If your team is building a dataset: How are we collecting ongoing feedback and responding to it?

8. Responsible coding terms

Advocate for and support replacing terminology in code that has harmful histories or connotations.

This includes terms that are derived from or associated with discrimination against particular groups, or which promote harmful stereotypes such as connecting blackness with good and whiteness with bad. There are commonly accepted replacement options — such as primary/secondary instead of master/slave. This recommendation applies to any software being developed.

Example: Google,[34] Apple,[35] Drupal,[36] and Linux[37] (among others) have started removing harmful terms from their coding platforms and using more inclusive language.

Questions to ask:

☐ Are there terms in our code with harmful origins, histories, or connotations?

☐ What alternatives can we replace these with, and how can we support advocacy in the broader tech community for more inclusive terms?

EGAL Tools:

- Terminology Guide: Harmful terms and alternatives tab - Use this to identify harmful terms in code and replacement options.

9. Language model limitations

If leading a team that is using an off-the-shelf language model to build an NLP tool, recognize the limitations, risks, and inherent issues and evaluate them in the context of the tool you are developing.

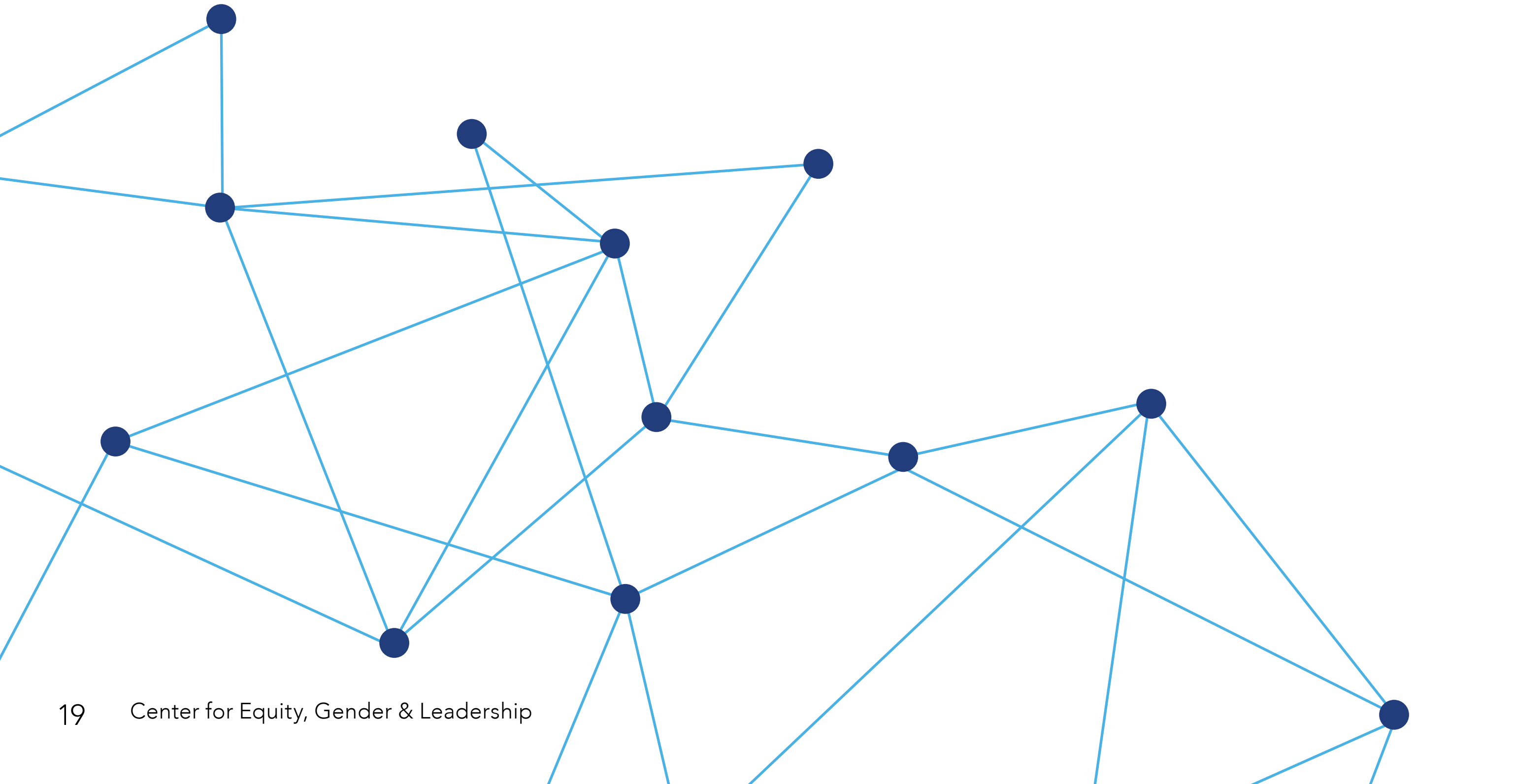
Increasingly, NLP tools are being developed using large language models (e.g., GPT-3, BERT). These language models serve as intermediary assets to adapt for various applications. They can be subject to the same limitations and harmful biases as the datasets discussed in #5. If adapting an off-the-shelf language model for your particular product it is important to understand whose voices (and therefore values, priorities, and perspectives) are represented in the model, and what the potential downstream risks might be for individuals, communities, and the organization. Use resources like PAIR's **Language Interpretability Tool** to help your team interact with language models and explore their shortcomings. Advocate for documentation, transparency, and explainability in off-the-shelf language models. As is possible, audit the off-the-shelf language model and carefully audit the NLP systems built using an off-the-shelf model. If building a language model that will inform NLP tools and be accessible for use by others, follow the practices outlined in #5.

Questions to ask:

□ Whose voices — and therefore values, priorities, and perspectives — are represented in the language model?

□ What are the potential downstream risks for individuals, communities, and our business?

□ How might we advocate for documentation, transparency and explainability in off-the-shelf language models? How might we audit the model we would like to use and the system we are building?



Call to action

The company that developed the AI system assessing English proficiency decided to scrap the existing tool and start over. From the start they prioritized the development of a tool that assessed English communication in an equitable and inclusive way. They also ensured their team reflected the diversity of potential users and followed the practices in this guide. Since launching their 2.0 tool, they have regained clients, drawn positive media attention for their efforts, and now benefit from high brand recognition and competitive advantage in the space.

Responsible AI innovation comes back to good leadership.

Advancing language for equity and inclusion in AI systems requires reimagining leadership priorities and asking: What is the purpose for which we are developing this AI system? Who is participating in design and development? Who benefits? Who may be harmed?

Advancing language for equity and inclusion in AI also requires creating an inclusive and supportive culture — one where concerns can be brought up, acknowledged, and worked through with empathy and understanding. Creating a more inclusive environment will require that those in power not only allow for marginalized voices to be heard, but also uplift and amplify those voices in positions of decision-making and power.

This critical work takes effort, time, and persistence. As leaders we are all on a journey to reflect, grow and build our own "equity fluency". Together, we can create the necessary changes to promote equity and ultimately enable a thriving and more equitable business world and society.

Appendix & Endnotes

Appendix 1. Methodology

This guide is informed by a systematic literature review that sought to 1) understand the linkages between language and power, 2) interrogate the linkages between ML and NLP in relation to inequality within business contexts, and 3) identify equitable and inclusive language. We searched academic journals (language / linguistics, social psychology, sociology, anthropology, ethnic studies) to inform these subjects. Beyond academic sources, we relied on other sources of literature including blogs, articles, reports, recordings, etc., and identified sources incorporating perspectives of community leaders related to language. We focused on race and the West, particularly the United States. Future research can do a deeper dive to explore how these practices apply globally across different identities. Finally, the search for sources was confined to the past 30 years (1990-2020), with the exception of seminal texts. In addition, we collected real-world examples related to challenges for management and technical teams in technology companies. Lastly, we received feedback from practitioners at a leading tech firm, and prototyped the guide with MBA students.

Appendix 2. Acknowledgments

The guide, along with the resources and tools, benefited from feedback and review by practitioners and leaders in the tech industry. The guide and resources also benefited from feedback and review by various individuals including Anna Gasparyan (Berkeley Haas), Antoni Lewa (Berkeley Haas), Chris Domina (IDEO), Cindy Tang (Berkeley Haas), Clemi Collett (University of Oxford), Faris Natour (UC Berkeley Human Rights Initiative & Article One Advisors), Gasper Begus (UC Berkeley), Jennifer Wells (Berkeley Haas EGAL), Marco Lindsey (Berkeley Haas), Ornella Tchoumie (UC Berkeley), Ruby Gao (Berkeley Haas), Sachin Bhattiprolu (Berkeley Haas), Seren Pendleton-Knoll (Center for Responsible Business at UC Berkeley), and Wendy Ake (Othering and Belonging Institute at UC Berkeley). We are grateful to those that prototyped the playbook with us and provided invaluable feedback including Elias Habbar-Baylac (Berkeley Haas), Hiro Kotegawa (Berkeley Haas), Jessica Hwang (Berkeley Haas), Madeline Deysine (Berkeley Haas), and Matt Solowan (Berkeley Haas).

Appendix 3. Glossary of key terms

African American English: The language varieties used by many Black Americans, though others may use terms including “African American Language,” “Black English,” “African American Vernacular English,” or “Ebonics.”[38] This variety of English is not more or less correct, expressive, or appropriate than any other variety. It has been systematically devalued as a result of racism and White supremacy, despite the empirical fact that it is linguistically equal to other language varieties including “Standard” American English.

Language: Language is a code used for communication. It includes spoken and signed languages, which are equal in their complexity and ability to communicate meaning. Both spoken and signed languages have grammatical structure and patterns that are fully formed and capable of expressing complex meaning.[39] Spoken and signed languages are learned naturally by children through exposure to language in their environments. All languages also show variation based on a variety of factors including context, age, race, gender, and region of origin. Languages are different from more fixed codes, such as computer languages or Morse code, in that meanings are not entirely predetermined; instead, people can use language to express themselves creatively and ambiguously.

Language models: Language models can be pre-trained and fine tuned for specific NLP tasks and downstream applications. They are systems trained to predict a sequence of words or characters based on preceding context or surrounding context.

Language variety: Language variety is a cover term used to describe all languages, dialects, and accents. Because the distinction between a language and a dialect or accent is arbitrary, we use "language variety" as a neutral term to describe any linguistic system.

Natural language processing (NLP): NLP is a sub-field of AI and ML at the intersection with linguistics. Its goal is to help computers process or understand "natural language" (language that has emerged naturally among and is used by humans) to perform human-like tasks.

"Standard" American English: A variety of English that is often used in media, politics, and education in the United States. It is based largely on the English used by middle-class White men. This variety of English is not more or less correct, expressive, or appropriate than any other variety; however, it has been accorded special status because of its association with people in power (who have historically tended to be White men).

Endnotes

[1] Fairness can have various definitions. In referencing "unfair" bias in AI systems, we mean bias that results in unjust impacts on people. When it comes to AI systems, justice considers how certain groups are oppressed or marginalized in the particular context and explores how the AI system can advance equity, rather than perpetuate a status quo that may oppress or marginalize certain groups.

[2] Gee, Gilbert C., Ro, Annie, Shariff-Marco, Salma, and Chae, David. (2009). Racial Discrimination and Health Among Asian Americans: Evidence, Assessment, and Directions for future research. Epidemiologic Reviews 31: 130-151. DOI: 10.1093/epirev/mxp009.

[3] Hanks, B. (2005). Pierre Bourdieu and the Practices of Language. Annual Review of Anthropology 34: 67-83. DOI: 10.1146/annurev.anthro.33.070203.143907

[4] Walley-Jean, J. C. (2009). Debunking the Myth of the "Angry Black Woman": An Exploration of Anger in Young African American Women. Black Women, Gender + Families, 3 (2): 68-86.

[5] Walley-Jean, J. C. (2009). Debunking the Myth of the "Angry Black Woman": An Exploration of Anger in Young African American Women. Black Women, Gender + Families, 3 (2): 68-86.

[6] Rhue, L. (2018). Racial Influence on Automated Perceptions of Emotions. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765.

[7] Hughes, C., and Mamiseishvili, K. (2014). Linguistic Profiling in the Workforce. In Marilyn Y. Byrd and Chaunda L. Scott (Eds.) Diversity in the Workforce: Current Issues and Emerging Trends (pp. 249-265). New York: Routledge.; Smalls, D. L. (2004). Linguistic Profiling and the Law. Stanford Law & Policy Review 15(2): 579-604.; Baugh, J. (2000). Racial Identification by Speech. American Speech 75(4): 362-364.; Baugh, J. (2016). Linguistic Profiling and Discrimination. In O. García, N. Flores, and M. Spotti (Eds.) The Oxford Handbook of Language and Society. 10.1093/oxfordhb/9780190212896.013.13.

[8] We adopt the use of the term "minoritized" in lieu of minority as a way to call out the power dynamics that lead some groups to be overrepresented or underrepresented in an organization or group (Sotto-Santiago, Sylk. 2019. Time to

Reconsider the Word Minority in Academic Medicine. J Best Pract Health Prof Divers 12(1): 72-78.). Using terms like "minoritized," which acknowledge that the inequities are being actively enforced, is more accurate than using passive terms like "minority" which hide that minoritization is a process that members of society carry out.

9 Bender, E., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of FAccT '21: 610-623. Doi: 10.1145/3442188.3445922.

10 Barthel, M., Stocking, G., Holcomb, J. & Mitchell, A. (2016). Reddit news users more likely to be male, young and digital in their news preferences. Pew Research Center. Retrieved from https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

11 Toxic twitter: A place for women. Amnesty International. Retrieved on March 9, 2021 from https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/.

12 Troll patrol findings. Amnesty International. Retrieved on March 9, 2021 from https://decoders.amnesty.org/projects/troll-patrol/findings.

13 Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. The Verge. Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

14 Abid, A., Farooqi, M. & Zou, J. (2021). Persistent anti-Muslim bias in large language models. Retrieved from https://arxiv.org/pdf/2101.05783v1.pdf#page=3.

15 Caliskan, A., Bryson, J. & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356 (6334): 183-186. DOI: 10.1126/science.aal4230

16 Hutson, M. (2017). Even AI can acquire biases against race and gender. Science. https://science.sciencemag.org/content/356/6334/183/tab-pdf.

17 "A Scalable Approach to Reducing Gender Bias in Google Translate." Google AI Blog, 22 Apr. 2020, ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html.

18 Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS, 115(16): 3635-3644.

19 Marcus, G. & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review. Retrieved from https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/.

20 Davidson, T., Bhattacharya, D. (2020). Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling. Retrieved from https://arxiv.org/abs/2005.13041.

21 Bohner, G. (2001). Writing about rape: Use of the passive voice and other distancing text features as an expression of pereived responsibility of the victim. British Journal of Social Psychology, 40: 515-529.

22 Blodgett, S. L. & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English

23 Tatman, R. (2017). Gender and dialect basis in YouTube's automatic captions. Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 53–59. Retrieved from https://www.aclweb.org/anthology/W17-1606.pdf.

24 (2018). Predictive tools across the hiring funnel. Upturn. Retrieved from https://www.upturn.org/reports/2018/hiring-algorithms/.

25 (2018). Predictive tools across the hiring funnel. Upturn. Retrieved from https://www.upturn.org/reports/2018/hiring-algorithms/.

26 Johnson, K. (2017, August 08). Stanford AI researchers make 'socially inclusive' NLP using Urban Dictionary and Twitter. Retrieved from https://venturebeat.com/2017/08/08/stanford-ai-researchers-make-socially-inclusive-nlp-using-urban-dictionary-and-twitter/

27 Project Euphonia. Google. Retrieved on September 23, 2021 from https://sites.research.google/euphonia/about/.

28 How Salesforce infuses ethics into its AI. Salesforce. Retrieved on September 30, 2021 from https://www.salesforce.com/news/stories/how-salesforce-infuses-ethics-into-its-ai/.

29 (2021). Responsible use of technology: The Microsoft case study. World Economic Forum. Retrieved from https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_2021.pdf.

30 (2020). Ethics by design: An organizational approach to responsible use of technology. World Economic Forum. Retrieved from https://www3.weforum.org/docs/WEF_Ethics_by_Design_2020.pdf.

31 Deighton, K. (2021). Tech firms train voice assistants to understand atypical speech. Wall Street Journal. Retrieved from https://www.wsj.com/articles/tech-firms-train-voice-assistants-to-understand-atypical-speech-11614186019.

32 Evaluating neural toxic degeneration. Allen Institute for AI. Retrieved on December 10, 2020 from https://toxicdegeneration.allenai.org/.

33 Langston, J. (2021). From conversation to code: Microsoft introduces its first product features powered by GPT-3. Microsoft AI Blog. Retrieved from https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/.

34 "Google Chrome and Android Move Away from 'Blacklist' - 9to5Google." Google, Google, 9to5google.com/2020/06/12/google-android-chrome-blacklist-blocklist-more-inclusive/.

35 Apple Style Guide. (n.d.). Retrieved from https://help.apple.com/applestyleguide/#/apdaf2bc3367

36 Remove usage of "blacklist", "whitelist", use better terms instead. (2021, January 14). Retrieved from https://www.drupal.org/project/drupal/issues/2993575

37 (n.d.). Retrieved from https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=49decddd39e5f6132ccd7d9fdc3d7c470b0061bb

38 King, S. (2020). From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans' Speech. Annual Review of Linguistics 6: 285-300. DOI: 10.1146/annurev-linguistics-011619-030556

39 "Sign Languages Are for Everyone!" (2020, September 23). United Nations. https://www.un.org/en/observances/sign-languages-day

egal

https://haas.berkeley.edu/equity/